

A COMPARATIVE EVALUATION OF FEATURE
SELECTION ALGORITHMS FOR CANCER
CLASSIFICATION THROUGH GENE EXPRESSION

DATA



ASLI TAŞÇI

DECEMBER 2016

A COMPARATIVE EVALUATION OF FEATURE
SELECTION ALGORITHMS FOR CANCER
CLASSIFICATION THROUGH GENE EXPRESSION
DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF
NATURAL AND APPLIED SCIENCES OF
IZMIR UNIVERSITY OF ECONOMICS

BY

ASLI TAŞÇI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

IN GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES

DECEMBER 2016

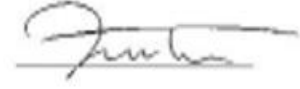
Approval of the Graduate School of Natural Science



Assoc. Prof. Dr. Devrim ÜNAY

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Engineering.

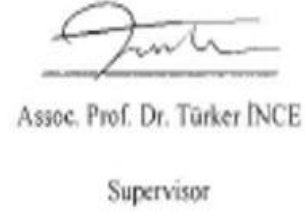


Assoc. Prof. Dr. Türker İNCE

This is the certify that we read this thesis and that in our opinion is in full adequate, in scope and quality, as a thesis for the degree of Master of Engineering.



Prof. Dr. Cüneyt GÜZELİŞ
Co-Supervisor



Assoc. Prof. Dr. Türker İNCE
Supervisor

Examining Committee

Prof. Dr. Murat AŞKAR

Prof. Dr. Cüneyt GÜZELİŞ

Assoc. Prof. Dr. Türker İNCE

Assoc. Prof. Dr. M. Alper SELVER

Assoc. Prof. Dr. Devrim ÜNAY



ABSTRACT
A COMPARATIVE EVALUATION OF FEATURE
SELECTION ALGORITHMS FOR CANCER
CLASSIFICATION THROUGH GENE EXPRESSION DATA

Aslı Taşçı

M.S. in Electrical and Electronics Engineering with Thesis

Graduate School of Natural and Applied Sciences

Supervisor: Assoc. Prof. Dr. Türker İnce

Co-Supervisor: Prof. Dr. Cüneyt Güzeliş

The number of people who have been diagnosed with cancer is increasing day by day. Cancer is diagnosed by interpreting the results obtained from the imaging technologies, blood analysis and diagnostic biopsies. Cancer begins in the cell. Therefore, studying genetic structure of the cancer cell is more reliable and informative in the long term. The analysis of the genetic structure of these cells can also be helpful while identifying marker genes, which can be used in targeted drug therapies. Additionally, understanding the gene networks, relations between genes and their products and the effects of genes on certain cell signaling pathways can help scientists to understand the dynamics of cancer. Microarrays are one of the important data sources for gene expression which can be used to diagnose cancer or classify cancer types. In this thesis, gene expression data from the benchmark datasets is analyzed to select a proper gene subset and classify three different types of cancer by using statistical and machine learning techniques. Nine different statistical filter approaches as feature selection methods are comparatively evaluated. For pattern recognition, support vector machines and multilayer perceptrons are employed to test the feature selection algorithms and classify cancer types. Keywords: Gene expression, cancer classification, gene selection, SVM, MLP

ÖZ

GEN İFADESİ VERİLERİ ARACILIĞIYLA KANSER
SINIFLANDIRMASINDA ÖZİNİTELİK SEÇME
ALGORİTMALARININ KARŞILAŞTIRMALI
DEĞERLENDİRİLMESİ

Aslı Taşçı

Elektrik Elektronik Mühendisliği Tezli, Yüksek Lisans

Fen Bilimleri Enstitüsü

Tez Danışmanı: Doç. Dr. Türker İnce

İkinci Tez Danışmanı: Prof. Dr. Cüneyt Güzeliş

Kanser teşhisi konan insanların sayısı her geçen gün artmaktadır. Doktorlar kanser türlerini, görüntüleme teknolojileri, kan analizi ve doku biyopsilerinden elde edilen sonuçları yorumlayarak teşhis ederler. Kanser hücrede başlar. Bu nedenle, kanser hücresinin genetik yapısının incelenmesi, uzun vadede daha güvenilir ve bilgilendiricidir. Ayrıca, bu hücrelerin genetik yapısının analizi, hedef ilaç tedavilerinde kullanılabilen belirteç genleri tanımlarken ve gen ağlarını, genler ile gen ürünleri arasındaki ilişkileri ve genlerin belirli hücre sinyal yolları üzerindeki etkilerini anlamakta da yardımcı olabilir. Mikro-dizilinler bu alandaki veri kaynaklarından biridir. Gen ifade değerlerini belirlerler ve kanseri teşhis etmek veya kanser türlerini sınıflandırmak için kullanılabilirler. Bu tezde önerilen yöntemde, gen ifadesi verileri, uygun bir gen alt kümesi bulmak ve kanser türlerini sınıflandırmak için istatistiksel teknikler ve makine öğrenme teknikleri kullanılarak analiz edilir. İstatistiksel filtre yaklaşımları, anlamlı bir gen alt kümesi elde etmek için öznelik seçme yöntemleri olarak kullanılır. Destek vektör makineleri ve çok katmanlı algılayıcı da öznelik seçme algoritmalarını test etmek ve kanser türlerini sınıflandırmak için kullanılır.

Anahtar Kelimeler: Gen ifadesi, kanser sınıflandırması, öznelik seçme, DVM

ACKNOWLEDGEMENT

I am very grateful to my supervisors, Assoc. Prof. Dr. Türker İnce and Prof. Dr. Cüneyt Güzeliş for their constant support on anything I came across during the research and writing process of this thesis.

I am grateful to my friends Seda Topuz and Erdem Okur for their precious support in every step of this thesis and my life.

Finally, I would like to dedicate this thesis to my beloved mother. I would not be here without her guidance and love.

TABLE OF CONTENTS

ABSTRACT	iii
ÖZ	iv
ACKNOWLEDGEMENT	v
LIST OF TABLES	viii
LIST OF FIGURES	xii
Chapter 1	1
Introduction	1
Chapter 2	5
Background	5
2.1 Cancer	5
2.2 Genetics, Epigenetics and Cancer	8
2.3 Microarray	10
2.4 Related Works	12
Chapter 3	17
Gene Selection and Classification Algorithms	17
3.1 Feature Selection Algorithms	17
3.1.1 T Test	18
3.1.2 Receiver Operating Curves	19
3.1.3 Bhattacharyya Distance	19
3.1.4 Wilcoxon Signed-Rank Test	20
3.1.5 Relative Entropy	20
3.1.6 ReliefF	21
3.1.7 Correlation based Feature Selection	21
3.1.8 Double Input Symmetrical Relevance	22
3.1.9 Maximum Relevancy Minimum Redundancy	22
3.2 Classification Algorithms	23
3.2.1 Support Vector Machines	23

3.2.2 Multi-Layer Perceptrons	27
Chapter 4	30
Experimental Results	30
4.1 Experimental Results for Leukemia Data	34
4.2 Experimental Results for Prostate Cancer Data.....	55
4.3 Experimental Results for DLBCL Data.....	77
Chapter 5	104
Conclusion	104
BIBLIOGRAPHY	107



LIST OF TABLES

- 2.1: Reference studies for leukemia dataset
- 2.2: Reference studies for prostate dataset
- 2.3: Reference studies for DLBCL dataset
- 4.1: Datasets' Description
- 4.2: Classification accuracy for leukemia data, No feature selection
- 4.3: Classification accuracy for prostate data, No feature selection
- 4.4: Classification accuracy for DLBCL data, No feature selection
- 4.5: Computational time of T test as feature selection algorithm
- 4.6: T test statistic feature selection for leukemia dataset (raw)
- 4.7: T test statistic feature selection for leukemia dataset (normalized)
- 4.8: Computational time of CFS as feature selection algorithm
- 4.9: CFS for leukemia dataset (raw)
- 4.10: CFS for leukemia dataset (normalized)
- 4.11: Computational time of Bhattacharyya Distance as feature selection algorithm
- 4.12: Bhattacharyya Distance for leukemia dataset (raw)
- 4.13: Bhattacharyya Distance for leukemia dataset (normalized)
- 4.14: Computational time of Entropy as feature selection algorithm
- 4.15: Entropy for leukemia dataset (raw)

- 4.16: Entropy for leukemia dataset (normalized)
- 4.17: Computational time of ReliefF as feature selection algorithm
- 4.18: ReliefF for leukemia dataset (raw)
- 4.19: ReliefF for leukemia dataset (normalized)
- 4.20: Computational time of Wilcoxon signed-rank test as feature selection algorithm
- 4.21: Wilcoxon signed-rank test for leukemia dataset (raw)
- 4.22: Wilcoxon signed-rank test for leukemia dataset (normalized)
- 4.23: Computational time of mRMR as feature selection algorithm
- 4.24: MRMR for leukemia dataset (raw)
- 4.25: MRMR for leukemia dataset (normalized)
- 4.26: Computational time of DISR as feature selection algorithm
- 4.27: DISR for leukemia dataset (raw)
- 4.28: DISR for leukemia dataset (normalized)
- 4.29: Computational time of ROC as feature selection algorithm
- 4.30: ROC for leukemia dataset (raw)
- 4.31: ROC for leukemia dataset (normalized)
- 4.32: Feature numbers for leukemia dataset using MLP (raw)
- 4.33: Feature numbers for leukemia dataset using MLP (normalized)
- 4.34: T test statistic feature selection for prostate dataset (raw)
- 4.35: T test statistic feature selection for prostate dataset (normalized)
- 4.36: CFS for prostate dataset (raw)
- 4.37: CFS for prostate dataset (normalized)
- 4.38: Bhattacharyya Distance for prostate dataset (raw)

- 4.39: Bhattacharyya Distance for prostate dataset (normalized)
- 4.40: Entropy for prostate dataset (raw)
- 4.41: Entropy for prostate dataset (normalized)
- 4.42: ReliefF for prostate dataset (raw)
- 4.43: ReliefF for prostate dataset (normalized)
- 4.44: Wilcoxon signed-rank test for prostate dataset (raw)
- 4.45: Wilcoxon signed-rank test for prostate dataset (normalized)
- 4.46: mRMR for prostate dataset (raw)
- 4.47: mRMR for prostate dataset (normalized)
- 4.48: DISR for prostate dataset (raw)
- 4.49: DISR for prostate dataset (normalized)
- 4.50: ROC for prostate dataset (raw)
- 4.51: ROC for prostate dataset (normalized)
- 4.52: Feature numbers for prostate dataset using MLP (raw)
- 4.53: Feature numbers for prostate dataset using MLP (normalized)
- 4.54: T test statistic feature selection for DLBCL dataset (raw)
- 4.55: T test statistic feature selection for DLBCL dataset (normalized)
- 4.56: CFS for DLBCL dataset (raw)
- 4.57: CFS for DLBCL dataset (normalized)
- 4.58: Bhattacharyya Distance for DLBCL dataset (raw)
- 4.59: Bhattacharyya Distance for DLBCL dataset (normalized)
- 4.60: Entropy for DLBCL dataset (raw)
- 4.61: Entropy for DLBCL dataset (normalized)

- 4.62: ReliefF for DLBCL dataset (raw)
- 4.63: ReliefF for DLBCL dataset (normalized)
- 4.64: Wilcoxon signed-rank test for DLBCL dataset (raw)
- 4.65: Wilcoxon signed-rank test for DLBCL dataset (normalized)
- 4.66: MRMR for DLBCL dataset (raw)
- 4.67: MRMR for DLBCL dataset (normalized)
- 4.68: DISR for DLBCL dataset (raw)
- 4.69: DISR for DLBCL dataset (normalized)
- 4.70: ROC for DLBCL dataset (raw)
- 4.71: ROC for DLBCL dataset (normalized)
- 4.72: Feature numbers for DLBCL dataset using MLP (raw)
- 4.73: Feature numbers for DLBCL dataset using MLP (normalized)
- 4.74: Most Commonly selected Genes for Leukemia dataset
- 4.75: Most Commonly Selected Genes for Prostate Dataset

LIST OF FIGURES

- 2.1: Microarray image
- 2.2: Hybridization and fluorescent dye labelling of microarray spots
- 3.1: Optimal Hyperplane for SVM
- 3.2 Linear Perceptron
- 3.3: Multi-layer Perceptron
- 4.1: Classification performance of SVM for leukemia data (raw)
- 4.2: Classification performance of SVM for leukemia data (normalized)
- 4.3: Classification performance of SVM for leukemia data (raw)
- 4.4: Classification performance of SVM for leukemia data (normalized)
- 4.5: Classification performance of SVM for leukemia data (raw)
- 4.6: Classification performance of SVM for leukemia data (normalized)
- 4.7: Classification performance of SVM for leukemia data (raw)
- 4.8: Classification performance of SVM for leukemia data (normalized)
- 4.9: Classification performance of SVM for leukemia data (raw)
- 4.10: Classification performance of SVM for leukemia data (normalized)
- 4.11: Classification performance of SVM for leukemia data (raw)
- 4.12: Classification performance of SVM for leukemia data (normalized)
- 4.13: Classification performance of SVM for leukemia data (raw)
- 4.14: Classification performance of SVM for leukemia data (normalized)

- 4.15: Classification performance of SVM for leukemia data (raw)
- 4.16: Classification performance of SVM for leukemia data (normalized)
- 4.17: Classification performance of SVM for leukemia data (raw)
- 4.18: Classification performance of SVM for leukemia data (normalized)
- 4.19: Classification performance of MLP for leukemia data (raw)
- 4.20: Classification performance of MLP for leukemia data (normalized)
- 4.21: Classification performance of SVM for prostate data (raw)
- 4.22: Classification performance of SVM for prostate data (normalized)
- 4.23: Classification performance of SVM for prostate data (raw)
- 4.24: Classification performance of SVM for prostate data (normalized)
- 4.25: Classification performance of SVM for prostate data (raw)
- 4.26: Classification performance of SVM for prostate data (normalized)
- 4.27: Classification performance of SVM for prostate data (raw)
- 4.28: Classification performance of SVM for prostate data (normalized)
- 4.29: Classification performance of SVM for prostate data (raw)
- 4.30: Classification performance of SVM for prostate data (normalized)
- 4.31: Classification performance of SVM for prostate data (raw)
- 4.32: Classification performance of SVM for prostate data (normalized)
- 4.33: Classification performance of SVM for prostate data (raw)
- 4.34: Classification performance of SVM for prostate data (normalized)
- 4.35: Classification performance of SVM for prostate data (raw)
- 4.36: Classification performance of SVM for prostate data (normalized)
- 4.37: Classification performance of SVM for prostate data (raw)

- 4.38: Classification performance of SVM for prostate data (normalized)
- 4.39: Classification performance of MLP for prostate data (raw)
- 4.40: Classification performance of MLP for prostate data (normalized)
- 4.41: Classification performance of SVM for DLBCL data (raw)
- 4.42: Classification performance of SVM for DLBCL data (normalized)
- 4.43: Classification performance of SVM for DLBCL data (raw)
- 4.44: Classification performance of SVM for DLBCL data (normalized)
- 4.45: Classification performance of SVM for DLBCL data (raw)
- 4.46: Classification performance of SVM for DLBCL data (normalized)
- 4.47: Classification performance of SVM for DLBCL data (raw)
- 4.48: Classification performance of SVM for DLBCL data (normalized)
- 4.49: Classification performance of SVM for DLBCL data (raw)
- 4.50: Classification performance of SVM for DLBCL data (normalized)
- 4.51: Classification performance of SVM for DLBCL data (raw)
- 4.52: Classification performance of SVM for DLBCL data (normalized)
- 4.53: Classification performance of SVM for DLBCL data (raw)
- 4.54: Classification performance of SVM for DLBCL data (normalized)
- 4.55: Classification performance of SVM for DLBCL data (raw)
- 4.56: Classification performance of SVM for DLBCL data (normalized)
- 4.57: Classification performance of SVM for DLBCL data (raw)
- 4.58: Classification performance of SVM for DLBCL data (normalized)
- 4.59: Classification performance of MLP for DLBCL data (raw)
- 4.60: Classification performance of MLP for DLBCL data (normalized)

4.61: Gene Expression Values for 24 commonly selected genes

4.62: Gene Expression Values for 25 commonly selected genes



Chapter 1

Introduction

Cancer is the disease of the era. According to the World Cancer Report issued by World Health Organization (WHO) in 2014 [1], there are approximately 14 million diagnosed cancer cases estimated and 8 million recorded deaths in 2012. These figures are foreseen to increase in the following decade by 42%, reaching up to 20 million cases in 2025. Most common types of cancer differ by sex. The leading forms of cancer in men are lung cancer, colorectal cancer, prostate cancer, liver cancer and stomach cancer. The most common cancers for women, on the other hand are breast cancer, lung cancer, cervix cancer, colorectal cancer and stomach cancer. Uncontrolled growth of cancer cases is caused by several factors. Aside from regional and ethnic factors of the mentioned disease, increase in obesity and tobacco usage, lack of physical activity and many different environmental or individual factors affect the spreading rate of the disease. Since cancer is not like any other traditional diseases, finding a permanent cure is more difficult. There is not one certain reason behind cancer like virus or bacteria. It does not target any specific organ. It can begin in one part of the body and easily spread to other parts. The behavior of the disease differs from patient to patient and can only be explained by studying the origin of the cancer. Cancer begins in the cell, and cell structure is unique to each individual. Therefore, there is not one specific drug, vaccine or treatment to cure cancer permanently for all cancer patients. Today, cancer growth can be controlled by several means of treatments such as chemotherapies, radiotherapies and immunotherapies. Even though continuous breakthroughs achieved by scientists, applied treatment methods are limited for cancer and are not easily accessible for everyone.

Cancer, as a disease, is the abnormal growth of cells and caused by the alterations in genetic or epigenetic structure of the cell. The main purpose of the studies on cancer treatment is to permanently repair the DNA damage caused by these alterations. Therefore, studying genetic structure of the cells in order to understand its behavior is of great importance to find a permanent cure for cancer, develop more effective drugs and vaccines for cancer treatments, diagnose the disease more accurately and make better prognosis predictions. Not only the structural changes in genes but also the gene interactions and gene networks are important to understand the development process of cancer. The improvements in the DNA sequencing technologies allow scientists to analyze excessive amounts of cellular data. Data mining and machine learning techniques are employed effectively to interpret the biological data. One of the widely used data types in terms of cellular data is gene expression values. Gene expression values can be measured at different levels of cellular processes. They basically give the information about how active the gene is.

An organism's life cycle begins with birth, later it grows and dies. Growth means cell division in the cellular level. The cell division process begins with the transcription of the related DNA sequence. Then, the transcribed sequence is translated to produce proteins and other types of cellular products. Improved techniques in this area allow scientists to measure different levels of this process. When measuring data in cellular level, there is a trade-off between biological relevancy and simplicity of the method of measurement. Since the DNA sequence is the source of the data, measuring is easy with the improved DNA sequencing methods. However; gene sequence is only the beginning of the process, how much and which type of information will be extracted from that sequence and how much of that information will be used to produce any type of cellular product are not clear. Therefore, biological relevancy increases, but the measurement methods become harder and cost more along with the specialty of the data. For instance, cell can transcribe excessive amount of messenger RNA (mRNA) but only use some of them to produce proteins. In this case, measurement in the proteomics level is more relevant in terms of biological accuracy, but it costs more than measurement process for the amount of produced mRNA. Microarrays are among the several types of measurement techniques for gene expression values. They cost relatively high but

provide excessive amount of biological data with regards to the activity of any specified gene by measuring the amount of produced mRNA. Microarrays, which serve a high throughput measurement method, have many advantages and disadvantages when it comes to data analysis. The major disadvantage of microarray data is the dimensionality. They usually have small sample size and excessive amount of gene expression values. This characteristic of the data makes it hard to be analyzed without employing pre-processing. Even though the biological relevancy is argumentative in this level of measurement, microarrays are used in many studies to classify cancer types, separate normal or cancerous tissues and identify marker genes for a specific type of cancer.

Alizadeh et al. used microarray of 4,026 genes to explore the subtypes of diffuse large B-cell lymphoma by using hierarchical clustering methods [2]. Begum et al. studied gene expression values to select the biomarker genes for leukemia using consistency based feature selection and k nearest neighbor (k NN), Naïve Bayes and Support Vector Machines (SVM) as classifiers to test selected genes' performances on classification of leukemia subtypes. Cystatin C and Nucleoside Diphosphate Kinase are selected as biomarker genes and the best classification performed by SVM with an average of 95% accuracy [3]. Khan et al. performed classification of small round blue-cell tumors (SRBCT) using hierarchical clustering and artificial neural networks [4]. Hu et al. used non parametric feature selection methods and compared the classification performance of classification trees and SVMs for seven different microarray datasets [5]. The aim of these studies is to help physicians make more accurate and fast diagnosis, use the biomarker gene information to develop targeted drugs for individuals and define an exclusive gene subset to better understand the dynamics of specific cancer types.

Feature selection is a vital part of microarray data analysis. Generally, microarray datasets for cancer classification involves high dimensional feature vectors with relatively smaller sample size. Genes are the features used to classify cancer in microarray datasets. Therefore, feature selection becomes selecting the most relevant genes for the cancer type and this may lead to biomarker gene identification which has an important role in many cancer related study areas. Non parametric tests, information theoretic approaches, probabilistic feature selection methods and genetic algorithms are practiced by many scientists to select the optimal

genes. There are several approaches to evaluate the suitability of the selected gene subsets. For instance, ranking algorithms may be used to rank the genes and top ranked genes are studied by the professionals in biology or medicine to assess the relation and importance of the genes. Another way of evaluating optimality of selected genes can be performed with an algorithm.

In this thesis, nine different feature selection algorithms are compared to select the most suitable gene subset for cancer classification. The suitability of the selected gene subsets is evaluated by using two different supervised machine learning techniques, Support Vector Machines (SVM) and Multilayer Perceptron (MLP). SVM's performance on big data is proven by many studies and often used for the analysis of microarray data. MLP are one of the most commonly used state of the art classifiers in machine learning area. Furthermore, different types of perceptrons (single/multilayer, linear/nonlinear) or artificial neural networks are employed by many studies [6] - [11]. The results of the comparison of proposed methods will provide better understanding about the practicality of feature selection and classification algorithms when studying gene expression.

The thesis is organized as follows: Chapter 2 provides comprehensive information about the biological background of the cancer dynamics and cellular processes. Chapter 3 reviews feature selection algorithms and define the criteria of nine different feature selection algorithms used in this thesis to select a suitable gene subset. Further, classification algorithms used to evaluate the classification performance of selected gene subsets are explained in Chapter 3. In Chapter 4, experimental results of three different microarray datasets with nine different feature selection and two classification algorithms are presented and explained in detail. Further a comparison of algorithms and biological relevancy discussions are provided in Chapter 4. Lastly, Chapter 5 includes final comments on the topic and future works.

Chapter 2

Background

2.1 Cancer

The simplest definition for cancer is uncontrolled cell growth. Each healthy cell usually undergoes a cell cycle, meaning the process of cell division. Normally, cells duplicate their DNAs, and this starts a series of incidents leading to cell division. If there is no DNA damage, cell divides and duplicates itself. In the presence of DNA damage, cell cycle gets interrupted. In this case, there are two options for a healthy cell. The first one is trying to fix the damage if it is possible. The second one is the programmed cell death (apoptosis). If the cell is untenable to fix itself, it usually goes to apoptosis. However; cancerous cells cannot perform this task. When there is damage in DNA that affects the cell cycle signaling pathways, the cell cannot follow the normal procedure and goes into proliferation (uncontrolled cell division). This abnormal cell growth may result in abnormal tissue growth which will eventually form a tumor in organs. In the case of leukemia, the damage occurs in the bone marrow cells which lead to an increase in the abnormal white blood cells. There are two types of tumors; benign and malignant. Since benign tumors do not spread to other tissues, they are not classified as cancerous tumors. On the other hand, malignant tumors have an invasive behavior and spread to other tissues and organs which make them cancerous.

Cancer is caused by the irreversible changes in DNA sequences or in the production of certain enzymes and RNA types (messenger RNA, micro RNA) whose changes will affect the cell cycle. There are several factors that can cause these

changes and affect cellular processes. Carcinogen substances, radiation, hormones, infectious diseases, diet and heredity are the main factors that may cause cancer [12].

Any substance that may cause cancer called carcinogens. Long term exposure to these substances can cause certain types of cancers. The most common forms of carcinogens are tobacco products. The most common cause of lung cancer is long-term usage of these products. There are certain chemicals that long term exposure to these chemicals that are proved to promote cancer after long-term exposure. For instance, asbestos fibers also cause lung cancer [13], [14].

Generally, the effect of radiation in cancer development is seen in the invasive cancer types such as skin cancer [15], [16].

Anything that affects the genes or cell cycle process has an important role in cancer development in cells. Therefore, hormones play an important role in cancer. Hormones that affect the cell cycle may cause the cell to proliferate and create the unwanted cancerous tissue. For instance, insulin, plays a key role in cell proliferation or estrogen level in the blood, has an effect on the development of breast cancer [17]-[19].

Some viruses, bacteria or parasites that cause infectious diseases may also be reason behind certain cancers. For instance, human papillomavirus (HPV) is an oncovirus (virus that has a role in cancer development in cells) which causes cervical cancer. Not as common as viruses, bacteria and parasites can cause cancer too [20], [21].

Unhealthy diet and lack of physical activity are the main reasons behind many diseases such as obesity, diabetes etc. Further, they have significant roles in nearly 10 types of cancers. There are researches that study the cancer metabolism and effects of sugar consumption, diabetes and insulin levels in the blood on tumorigenesis [22]-[24].

Despite the fact that only 10% of cancer is caused hereditarily, the effect of inheritance cannot be ignored. Since cancer is caused by the mutations in the DNA, inheriting these corrupted DNAs increase the risk of cancer [25]-[27]. Therefore, studying genomics and epigenetics has a significant role in cancer research. It can help in all the three stages of cancer: diagnosis, treatment and prognosis.

Most of the cancer types are difficult to be diagnosed. The abnormal cell growth is not easy to be detected because it takes time for a tumor to form, and symptoms might not appear until the last stages of cancer. Therefore, regular checkups play an important role in the diagnosis process of cancer. Even after the checkups, for an accurate diagnosis of cancer, the cancerous tissue must be examined by a pathologist.

Even though there is no permanent cure for cancer, there are several treatment options [28]. Treatment method can change depending on the type and location of the cancer. The most common method is chemotherapy which is basically using drugs to kill the cancerous cells. Most types of cancer respond to chemotherapy but it is toxic to the body and has many side effects. Therefore, it has a limited usage but it can shrink the tumors and reduce some of the symptoms.

Another way to treat cancer is the radiation therapy. Ionizing radiation is used to damage the DNA of cancer cells and destroy them. It is a more intense method than chemotherapy since focused radiation beams directly affect the tumor [29]. Generally, chemotherapy and radiation therapy used along with the surgery. For cancers that form a solid tumor, surgery can be a permanent cure. In most of the cases, the tumor gets shrunk by using chemo or radiation therapy then removed surgically. Even though the cancerous tissue is completely removed, there is no guarantee that cancer will not relapse. Recently, there are many researches on immunotherapy to cure cancer. Immunotherapy is the activation of the immune system to fight the cancer by using antibodies but this treatment method is still experimental.

Another vital part of cancer treatment is prognosis, which is used to determine the life span of cancer patient. Cancer types, the invasive behavior of the cancer, mental and physical health of the patient are the main factors that affect the survival term of the patient. The TNM (Tumor Node Metastasis) staging system helps physicians to assess the cancer development in the body, determine a treatment and make proper prognosis estimations. It is only applicable to the cancers caused by solid tumors. The system focuses on three main aspects of a tumor; the size of the tumor, numbers of the affected lymph nodes by the tumor and metastasis.

Cancer has a heterogeneous structure. It can be caused by chromosomal changes, genetic or epigenetic alterations, environmental or hereditary factors etc. Even though the DNA sequencing techniques provide high-throughput data about the genetic and epigenetic factors of cancer, it is not possible to identify one source for cancer development. Therefore, physicians have to consider all these aspects of the disease when they administer treatments and make prognosis estimations. The following sub section will provide a detailed background on the cellular mechanisms which play a role in cancer dynamics.

2.2 Genetics, Epigenetics and Cancer

Cancer is the general name for the disruptions in the cell growth mechanisms. The natural process for a cell is to grow and die. In order to accomplish this, cells must complete a process called “cell cycle”. This procedure includes different phases; each of them is regulated by certain types of proteins and enzymes. There are several cell cycle checkpoints to control the regulation of the cell division procedure. If there is DNA damage in the cell, it is detected in these checkpoints and cell division procedure gets interrupted. If the cell is able to repair the damage, it continues to divide and grow. If not, the cell goes to apoptosis. For a cell to become cancerous, genetic changes have to be in the genes that control the cell growth and division. The types of genes, whose changes may cause cancer, are oncogenes and tumor suppressor genes. Oncogenes usually have regulatory functions in the cell cycle, and changes in these genes may occur in several ways during the cell division. The cell may encounter a change in the chromosomal level or in the nucleotide sequence of the gene. These nucleotide sequence changes are called mutations. When a mutation occurs in an oncogene, it prevents cell from dying, and cell goes into proliferation. Proliferated cells form cancerous tissue which will provoke tumorigenesis. For instance, MYC (V-Myc Avian Myelocytomatosis Viral Oncogene Homolog)¹ gene is an example of oncogene. This gene takes part in the transcription of DNA during the cell division and if there is any mutation or overexpression present in this gene, cell cycle progress is interfered. Tumor suppressor genes have regulatory functions and protect cells from any cancerous development. TP53 (tumor

¹ National Center for Biotechnology Information, 2016
<https://www.ncbi.nlm.nih.gov/gene/4609>

protein p53)² is an example of a tumor suppressor gene; alterations in this gene can be seen in more than 50% of the cancer types. Even though alterations in these types of genes are important, they are not the only responsible factors of a cell's cancerous behavior [30], [31].

Epigenetics is the study of changes which affect a gene's output. A gene's output can be a protein or RNA such as messenger RNA (mRNA), transfer RNA (tRNA), micro RNA (miRNA) etc. Epigenetic changes do not occur in the DNA sequence of a gene but may affect a gene's activation or expression. DNA methylation and histone modification are two main factors for epigenetic changes. DNA methylation is the addition of a methyl group to the DNA which cause disruptions in the function and expression of that gene. Histones are proteins that have a part in DNA structure. DNA sequences comprised of base pairs are meters long and could not fit in a cell if they do not form into nucleosomes. DNA sequences wrap around histone proteins and form nucleosomes which are the constituent for chromatins. Therefore, they are able to fit in a cell or cell nucleus. Any alterations in these histone proteins affect the expression of a gene as well. Gene expression is the process of a protein or RNA production. Mainly, this process includes transcription and translation of a gene. In transcription, DNA strands are separated, and coding region of a particular gene gets duplicated into types of RNAs. In case of protein coding genes, transcribed RNAs are mRNAs, and they produce proteins with the translation process. If the gene is a non-coding gene, several types of RNA can be transcribed and become a mature RNA as the result of transcription. The improvements in DNA sequencing technologies made interpreting gene expression values easier. There are several methods to measure gene expression values. The oldest method to measure gene expression is differential display which is based on the comparison of two RNA strands. Although it is easy to implement, gene expression values measured by differential display can be unreliable and sensitive to errors. Northern blotting is another outdated method for gene expression measurement. The main idea of Northern blotting is to measure the excess amount of RNA by using radioactive probes which make the technology undesirable despite being cheap. Real time and reverse transcriptase polymerase chain reactions (RT-

²National Center for Biotechnology Information, 2016
<https://www.ncbi.nlm.nih.gov/gene/7157>

PCR) are other methods for measuring gene expression. The basic principle for these two methods is the same. They measure the output of polymerase chain reaction which produces specific DNA or RNA sequence in large quantities. Fluorescence dyes are used to label sequences. RT-PCR is a reliable method to measure gene expressions, but its application is costly and the size of output data is very limited [32], [33]. The most recent methods used to measure gene expression values are Serial Analysis of Gene Expression (SAGE) and Microarray. SAGE is the analysis of the sequenced, tagged and amplified complementary DNA samples produced by mRNA samples. Despite being a reliable method, it is very expensive to conduct [34]. Microarray is the measurement method used to produce the datasets studied in this thesis. Therefore, a detailed explanation of the method is provided in the next subsection.

2.3 Microarray

Microarray is the most common measurement method for gene expression values. They provide information about thousands of genes in a single experiment. Microarrays are basically glass slides with thousands of spots on them. These spots contain DNA strands or sequences which represent a gene. Microarrays are the clean way of Northern blotting. The basic principle of Northern blotting was measuring the excess amount of mRNAs by radioactive labelling but in microarrays instead of radioactive labels, fluorescence dyes are used for labelling the data. Microarray experiments have several significant processes and always have the same processes for a reference sample from the same cell/tissue type to see the difference in the analyzed sample. First, mRNA samples should be separated from the cells. This cell could belong to a tumor, blood sample, bone marrow or lymph node depending on the cancer which will be analyzed. After the mRNAs extracted from the sample, reverse transcription of the mRNAs begin to produce complementary DNA strands (cDNA) and these cDNA sequences are labelled with fluorescence dyes. Then, complementary DNA strands washed to the microarray slide which is the process of cDNA sequences attaching to their complementary sequence on the slide. This process is called hybridization. After the samples hybridized to the slide, they are agitated with laser and scanned with a laser scanner. The more the cDNA strands get hybridized and bound to the spots, the more agitated they get and radiate or in the

case of fluorescence labelling emit more fluorescence light. Laser scanner measures the amount of emitted fluorescence lights by the microarray slides.

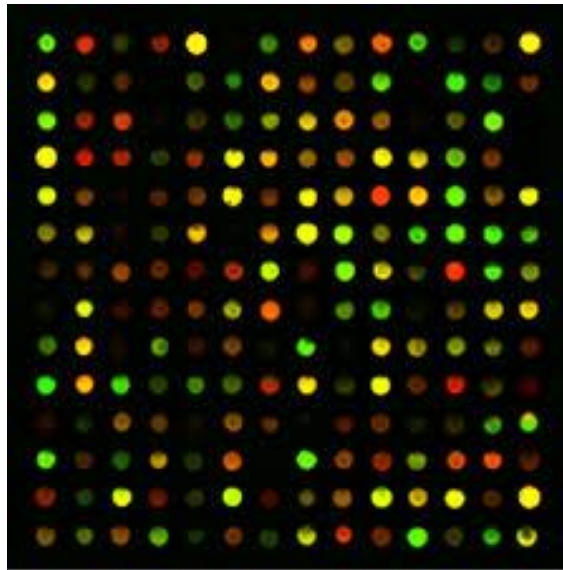


Figure 2.1: Microarray image

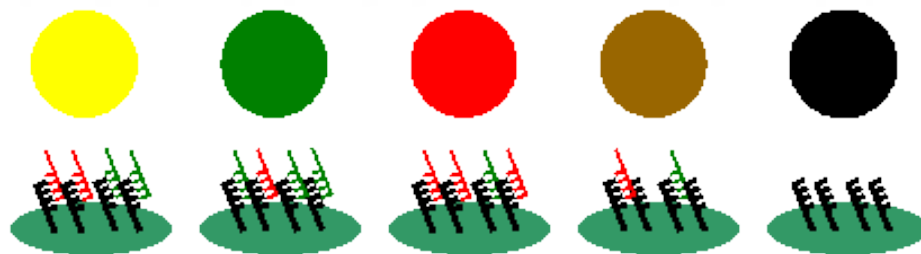


Figure 2.2: Hybridization and fluorescent dye labeling of microarray spots

An image of the dyed microarray slide is produced after the hybridization and scanning processes. Intensity values of the spots are detected with advanced image processing techniques. The first and most important image analysis part is determining the spots. According to the expression levels, spots sizes may change as shown in Figure 1. If the gene expression value is low, it might appear obscure on the microarray image. Therefore, identifying spots on the microarray image is a vital part of the process. The next important part is the intensity analysis. Intensity analysis of the spot and background gives the gene expression value. There are two

approaches for the intensity analysis of the spots. One approach measures the intensity for the spot signal and background by accepting fixed spot sizes. This approach is easy and cheap but has a great error potential. The other approach is to determine exact spot sizes and analyze the intensity values in defined ranges for the spots. This method result in more accurate expression values but expensive and computationally challenging. Lastly, normalization is performed for the image to highlight the differentially expressed genes [35]. For further details on the microarray experiments and image processing, one may refer to [36]-[39].

2.4 Related Works

Cancer studies have achieved rapid improvements over the past two decades. Thanks to discoveries on the dynamics of cancer and biology, scientists have been able to take control of the disease, diagnose more accurately and estimate better prognosis for the patients. Naturally, these rapid discoveries have not been the result of only biology or medicine research areas. Several approaches by scientists from different areas are used to analyze the dynamics of the disease. Recently, with the improvements in DNA sequencing technologies, scientists have been able to study excessive amounts of cellular level data. When biological data became interpretable for computers, scientists started using statistical approaches more effectively. These improvements reduced the time of the analysis and made the topic open to scientists from a variety of areas. Microarray is one of the data types used by statisticians and physicians to classify cancer types and identify marker genes for diagnostic or prognostic purposes. For the microarray data, dimensionality is a big problem when it comes to analysis of the data. Sample numbers are limited and feature numbers are too high. Therefore, statistical approaches are biased and do not provide good results without any pre-processing. Feature selection is important in order to analyze the data. Further, a meaningful and optimal feature subset should be provided to distinguish relevant genes with the cancer type. Firstly, simple and strict filter approaches were used to reduce the feature number. For instance, any gene expression value that is above or below the pre-defined limits was cut-off. This type of filters does not take the in-between feature relations or feature to class relations into consideration, which may result in biologically irrelevant and redundant features. Later on, scientists started to consider these gene relations and use more effective

filter approaches to find better feature/gene subsets. Golub et al. provided a microarray dataset for the classification of two different kinds of leukemia. A weighted voting scheme is used as class predictor, and “prediction strength” is calculated for each gene by using the following equation;

$$PS = \left| \frac{\mu_{ALL} - \mu_{AML}}{\sigma_{ALL} - \sigma_{AML}} \right| \quad (2.1)$$

A preset threshold of 0.3 was used to select the “informative genes” and classify the data using these genes. A set of 38 samples were used as training data and another set of 34 samples were used as test set. Class predictor was able to correctly classify 29 samples of the total test set [40]. Golub’s dataset is made publicly available and used for many other studies with a variety of classifiers and feature selection algorithms.

Statnikov et al. used support vector machines, k -nearest neighbors, backpropagation neural networks and probabilistic neural networks as classifiers for eleven different microarray datasets which include binary or multi-class cancer classification problems. Before any type of analysis, normalization so that the data will have 0 mean, 1 standard deviation and scaling into the range of [0 1] is performed. Three different feature selection methods used to rank the features; ratio of genes between-categories to within-category sums of squares, Signal to noise ratio and Kruskal-Wallis non-parametric one-way ANOVA. Different number of genes was used for classification to see how the number of genes will affect the classification performance. The best classification performance for leukemia data is achieved by using SVM with the accuracy of 97.5% [41].

Lee et al. used many different classifiers and feature selection algorithms on seven different microarray datasets which include binary or multi-class cancer classification problems. Several pre-processing steps are performed on the datasets. For the leukemia data, a floor of 100 and ceiling of 16,000 thresholding, filtering of the genes according to their maximum and minimum values ratio ($\frac{\max}{\min} \leq 5$ or $(\max - \min) \leq 500$) and base 10 logarithmic transformation is applied to the data. Consistent

with the previous studies, SVM gave the best result with 94% accuracy for the classification of leukemia types [42].

Furey et al. used SVM as classifier with the Golub's prediction strength as feature selection method and applied this model to an ovarian cancer dataset. Further, for the sake of completeness, the method is verified on two previously published datasets. SVM was able to classify the leukemia types with 91% accuracy [43].

In another study by Guyon et al., 100% classification is achieved for the leukemia dataset using recursive feature elimination method with SVM [44].

Table 2.1: Reference studies for leukemia dataset

Reference Study	Pre-processing	Gene Selection Method	Classifier	Classification Accuracy
(Alexander Statnikov, 2005)	Normalization & Scaling to [0 1]	BW Signal to noise ratio ANOVA	SVM <i>k</i> NN NN PNN	97.5% 83% 76% 85%
(Jae Won Lee, 2005)	Thresholding, Filtering, Log transformation, Normalization & Scaling	BSS/WSS Wilcoxon Soft - Thresholding	FLDA DLDA DQDA Logistic regression <i>k</i> NN CART SLNN ML NN SVM	78% 88% 87% 78% 90% 85% 92% 79% 94%
(Terrence S. Furey, 2000)	Normalization	Prediction strength	SVM	91%
(Isabelle Guyon, 2002)	Normalization	SVM-RFE	SVM	100%

Another microarray dataset for normal tissue and cancerous tissue classification of prostate cancer was published and analyzed by Singh et al. [45]. Normalization so that the data will have 0 mean and 1 standard deviation, thresholding in the limits of 100 and 16,000 and a variance filter which filters out the genes whose expressions does not vary more than 5-fold between two samples applied as pre-processing to the data. *K*-nearest neighbor was used as classifier with signal to noise ratio as feature selection method. *k*NN was able to classify the data

with the accuracy ranging from 86% to 92% which was not recorded as an acceptable rate by Singh et al. but the selected genes used for the further biomarker gene studies.

Statnikov et al. studied prostate cancer dataset as well. The same pre-processing and gene selection methods were used for the prostate data as for the leukemia data. The best classification performance was achieved by SVM with the accuracy of 92% [41].

Peng et al. studied the prostate cancer dataset without any pre-processing and random feature selection. Different types of SVMs used as classifiers and ensemble SVM gave the best classification result with the accuracy of 95.1% [46].

Table 2.2: Reference studies for prostate dataset

Reference Study	Pre-processing	Gene Selection Method	Classifier	Classification Accuracy
(Dinest Singh, March 2002)	Normalization, Thresholding & Variance Filter	Signal to Noise Ratio	kNN	86-92%
(Alexander Statnikov, 2005)	Normalization & Scaling to [0 1]	BW Signal to noise ratio ANOVA	SVM kNN NN PNN	92% 85.09% 79.18% 79.18%
(Peng, 2006)	No pre-processing	Random gene selection	SVM (all genes) SVM(Random) Bagging(all genes) Bagging (Random) Boosting (all genes) Boosting (Random) enSVM (Random)	91.2% 92.2% 91.2% 90.2% 89.2% 89.2% 95.1%

The last dataset is for the classification task of diffuse large B-cell lymphoma and follicular lymphoma. Shipp et al. published and analyzed the data by using signal to noise ratio as gene selection algorithm and a weighted voting scheme as classifier. Normalization, scaling, thresholding and variance filter were applied to the data before classification. Thresholding applied between the range of 20 and 16000. Expression values lower than 20 were fixed to 20 expression values and the ones

greater than 16000 were fixed to 16000 gene expression values. Further, if gene's expression did not vary, they are filtered out. Classification was performed with 30 gene predictors and 91% accuracy was achieved [47].

Pankaj et al. used different three different gene pairing methods and six different classifiers. Genes were paired, and gene expression vectors were calculated according to sum, difference, multiplication or signs of the genes. The best classification performance was achieved by Top Scoring Pair with 98.10% accuracy [48].

Kun Yang et al. proposed two new gene scoring methods: GS1 and GS2. F-test and Cho's method were used additional to these proposed methods as gene selection algorithms. SVM and k NN were used for the classification part of the problem. Intensity thresholding between 20 and 16000 gene expression units and filtering according to the maximum and minimum value relations of the genes ($\frac{\max}{\min} \leq 3$ or $(\max - \min) \leq 100$) were applied before the classification. The best result for the DLBCL data was achieved by GS2-SVM by using 5-fold cross validation. When Leave-one-out was used for cross validation, and classification performance raised to 96% [49].

Table 2.3: Reference studies for DLBCL dataset

Reference Study	Pre-processing	Gene Selection Method	Classifier	Classification Accuracy
(Margaret A. Shipp, January 2002)	Normalization,, Thresholding & Variance Filter	Signal to Noise Ratio	Weighted Voting Classification	86-92%
(Pankaj Chopra, 2010)	-No pre-processing	Gene pairing	PAM SVM k NN DT TSP k -TSP	85.45% 97.40% 89.61% 80.52% 98.10% 97.40%
(Kun Yang, 2006)	Thresholding & Filtering	GS1 GS2 Cho's F test	k NN SVM	92% 93%

Chapter 3

Gene Selection and Classification Algorithms

Gene expression values have been a great source of information to understand cellular mechanisms. Sequencing technologies enabled to quantify these values and made biological information to be interpretable for computers. DNA microarrays are one of the methods for measuring gene expression values. Microarrays provide high throughput gene expression data, but this extensive information has many disadvantages when it comes to analyzing it for various purposes. One of them is the dimensionality problem when performing a classification. Even though a classifier biased to the dimensionality problem is used; a suitable gene subset should be selected to provide relevant genes with the classified disease. The relevancy of the selected gene subsets can be assessed through many approaches. One approach is to apply only feature selection or ranking algorithms to the genes and examine the top ranked or selected genes by directly analyzing the biological functions of them. Another approach is to use statistical or machine learning techniques to evaluate the classification ability and importance of the selected genes [50]. In this thesis, nine different feature selection algorithms are used to select the suitable gene subset and two different classifiers are used to evaluate the importance of the selected genes subsets. The next subsection will provide a comprehensive explanation of these algorithms.

3.1 Feature Selection Algorithms

Feature selection plays a crucial part in the analysis of microarrays and cancer classification. Selected genes are expected to provide information about the

characteristics of cancer. Despite achieving a good classification accuracy is important, the relevancy of the selected genes is vital for the analysis as well. There are two types of feature selection algorithms; filter and wrapper approach. Filter methods only consider feature relations and analyze them with statistical approaches such as correlation, mutual information etc. They do not depend on the classifier model, and they are faster. Yet, some of the selected features might be redundant and need post-processing to select the best feature subset. Wrapper methods depend on the classifier model. Since they test each possible subset with a learning algorithm to find the best one, they need more computing time, and they perform very slowly with big datasets. Further, they are prone to overfitting if observation number is low. Therefore, filter methods are more suitable for microarray datasets because of their small sample size and large number of features.

In this thesis, nine different filter type feature selection algorithms are used to select a suitable gene subset for three different microarray datasets, and classification is performed to distinguish cancerous tissues from normal samples or determine the cancer type. Further, feature selection algorithms' performance is tested according to their provided gene subset's classification performance. These feature selection algorithms are t test, Receiver Operating Characteristics curves (ROC), Bhattacharyya distance, Wilcoxon signed-rank test, Entropy, ReliefF algorithm, Correlation Based Feature Selection (CFS), Double Input Symmetrical Relevance (DISR) and Maximum Relevancy Minimum Redundancy (mRMR). While the remaining algorithms rank the features according to defined criteria, mRMR and DISR methods provide a specified gene subset by previously defined number.

3.1.1 T Test

T test is the test statistic to determine differences between two sets of data. Generally, t statistic gives the best result for the data pairs which have normal distribution and independent from each other [51], [52]. The difference between the data is calculated as follows:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \quad (3.1)$$

where μ_1 and μ_2 are the estimates for the mean values, s_1 and s_2 are the estimates for the variances, N_1 and N_2 are the sample numbers of class 1 and class 2, respectively. In this thesis, absolute value of the t-test with pooled variance estimate used as ranking criterion for the feature selection from three different microarray datasets. The assumption in the pooled variance estimation is that the samples have equal variances and calculated by the following formula:

$$s_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \quad (3.2)$$

Then the t value for the t -statistic is calculated as:

$$t = \frac{\mu_1 - \mu_2}{s_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \quad (3.3)$$

3.1.2 Receiver Operating Curves

Receiver operating curves are generally used to analyze the performance of a classifier. They are plotted according to false positive and true positive rates of a classifier which describe the sensitivity and specificity of it. In this thesis, the intersection area of ROC and random classifier slope is considered as a ranking criterion for the features [53]-[55].

3.1.3 Bhattacharyya Distance

Bhattacharyya distance is a divergence measure. It provides the degree of similarity between two variables. There are two approaches for the Bhattacharyya distance. One is by calculating the Bhattacharyya coefficient and using the following formula to calculate Bhattacharyya distance.

$$B(P_1, P_2) = -\ln(\rho(P_1, P_2)) \quad (3.4)$$

where P_1 and P_2 is the probability distributions for the two distinct samples in the same feature domain. $\rho(P_1, P_2)$ is the Bhattacharyya coefficient and calculated by the following formula:

$$\rho(P_1, P_2) = \sum_{f \in F} \sqrt{p_1(f)p_2(f)} \quad (3.5)$$

where f is a feature in the feature domain F . In this thesis, Bhattacharyya distance is calculated by minimizing the minimum attainable classification error [56].

3.1.4 Wilcoxon Signed-Rank Test

Wilcoxon signed – rank test is a non-parametric feature ranking test similar to t-test statistic. Unlike the t-test, features are not considered to have normal distribution. This characteristic of the method makes it preferable for the data which have outliers. Features are ranked according to the difference between pairs. Then, a W statistic is calculated for the feature pairs and used to evaluate the features [51], [57]. W value is defined by the following equation:

$$W = \sum_{i=1}^{N_f} [\text{sign}(x_{2,i} - x_{1,i})R_i] \quad (3.6)$$

where N_f is the number of feature pairs, R_i is the rank for the i^{th} feature. In this thesis, a specific form of Wilcoxon signed – rank test is used to rank feature. Mann–Whitney U test is the equivalent for two sample t test with pooled variance estimation; it assumes equal variance for the samples but does not require samples to have normal distribution. In Mann – Whitney U test, features have assigned ranks and U statistic is calculated to see the relation of the feature to the class and importance of the features is evaluated according to U statistic. In this thesis, U statistic is the determined criteria for feature selection and calculated by the following equation:

$$U = N_1N_2 + \frac{N_2(N_2 + 1)}{2} - \sum_{i=N_1+1}^{N_2} R_i \quad (3.7)$$

3.1.5 Relative Entropy

In information theory, entropy defined as a purity measurement for the transmitted signals. It basically represents the amount of information which is

expected to specify the quality and sufficiency of the information [58], [59]. Entropy is calculated by the following formula:

$$H(X) = -\sum_{i=1}^N P(x_i) \log_b P(x_i) \quad (3.8)$$

In this thesis, relative entropy is used as the ranking criterion for features. Relative entropy in another words information gain is defined by the following equation:

$$RE(X) = H(X) - H(X/Y) \quad (3.9)$$

where X and Y are the features, H(X/Y) is the conditional entropy of X.

$$H(X/Y) = -\sum_{y \in Y} P(y) \sum_{x \in X} P\left(\frac{x}{y}\right) \log_b P(x/y) \quad (3.10)$$

3.1.6 ReliefF

Relief is a weight-based feature selection algorithm. When ranking features according to the Relief algorithm, samples are divided into the classes and Near-hit, Near-miss values are calculated for a randomly chosen feature. Then, weight of the feature is updated using the following equation:

$$W_i = W_i - \text{diff}(x_i, \text{near-hit}_i)^2 + \text{diff}(x_i, \text{near-miss}_i)^2 \quad (3.11)$$

Relief algorithm uses nearest neighbor approach to find the near-hit sample which is the closest sample from the same class and near-miss sample which is the closest sample from the different class. In this thesis, an updated version of Relief algorithm called ReliefF is used to rank the features according to the weights calculated by using this algorithm. ReliefF algorithm is more accurate and applicable to the incomplete or multi class data version of the Relief algorithm. ReliefF algorithm finds k nearest neighbors to calculate the near-hit and near-miss values for the feature and makes an estimation using the average information k-nearest neighbors [60]-[64].

3.1.7 Correlation based Feature Selection

CFS algorithm proposed by Hall et.al. provides a feature subset according to their correlation in between features and to the class [65]. In this study, feature

number for the gene subset is accepted equal to the complete number of features and calculated CFS score for each is used to rank the features/genes. Correlation score for each gene is evaluated individually but considering the in between feature correlation reduces the redundancy of the selected genes.

$$CFS = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \quad (3.12)$$

CFS is the score for ranking, \bar{r}_{cf} is the average correlation of feature to class, \bar{r}_{ff} is the average correlation between features and k is the number of features in the gene subset which in this case, is equal to the total number of features [63].

3.1.8 Double Input Symmetrical Relevance

Double input symmetrical relevance feature selection method uses the mutual information and entropy measures to determine the relevancy of the feature to the class [66], [67]. First, a symmetrical relevance measure is calculated by using the following equation:

$$SR(X_S; Y) = \frac{I(X_S, Y)}{H(X_S, Y)} \quad (3.13)$$

where $I(X_S, Y)$ is the mutual information between feature subset S and class Y, $H(X_S, Y)$ is the entropy of feature subset S and class Y. The best feature subset is defined by the maximizing the symmetrical relevance of the features calculated by the following criterion:

$$X_S^{DISR} = \arg \max \sum_{X_i \in X_S} \sum_{X_j \in X_S} SR(X_{i,j}, Y) \quad (3.14)$$

where $X_S \in X$.

3.1.9 Maximum Relevancy Minimum Redundancy

The best feature is the most relevant and the least redundant feature and this algorithm chooses the best features by analyzing information theoretic relations between features and class [66]-[68]. The relevancy of the feature is defined by the following equation:

$$D(S, c) = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \quad (3.15)$$

where S is the feature subset and $I(x_i, c)$ is the mutual information of the i^{th} feature to the class. Even though maximum related features are selected for the classification, there can be redundant features in this most relevant features subset. Therefore, redundancy of the features is measured by the following criterion:

$$R(S) = \frac{1}{|S|} \sum_{x_i, x_j \in S} I(x_i, x_j) \quad (3.16)$$

The mRMR algorithm aims to find the best balance between relevancy and redundancy. This is achieved by maximizing $\text{Max } \Phi(D, R)$ where $\Phi(D, R)$ is defined by the following criterion:

$$\Phi = D - R \quad (3.17)$$

The performance of the feature selection algorithms are tested by using SVM and MLP. Therefore, these classifiers will be explained in the next section.

3.2 Classification Algorithms

Cancer studies using gene expression data have gained great importance over the past two decades. Scientist used several statistical or machine learning approaches to classify cancer types and define gene subsets or biomarker genes. In this thesis, importance of the selected genes is evaluated by using two classification algorithms. Support vector machines and multilayer perceptrons are used for the classification of three different microarray datasets all of which consist of binary classification problems for different cancer types. Thus, detailed information about the classifiers is given in the next subsections.

3.2.1 Support Vector Machines

Support vector machines are supervised classification or regression algorithms, and widely used in the text, image, handwriting recognition and bioinformatics. First, SVM is proposed for binary classification problems but multi class adaptation is provided later. One-versus-one (OVO) or one-versus-all (OVA) approaches are used for multi class classification problems. First, SVMs were

applicable to the linear and separable data. Later on, SVM is improved for nonlinear and non-separable data by using kernel functions. In the case of linear and separable data, a hyperplane is defined between samples to separate two distinct classes.

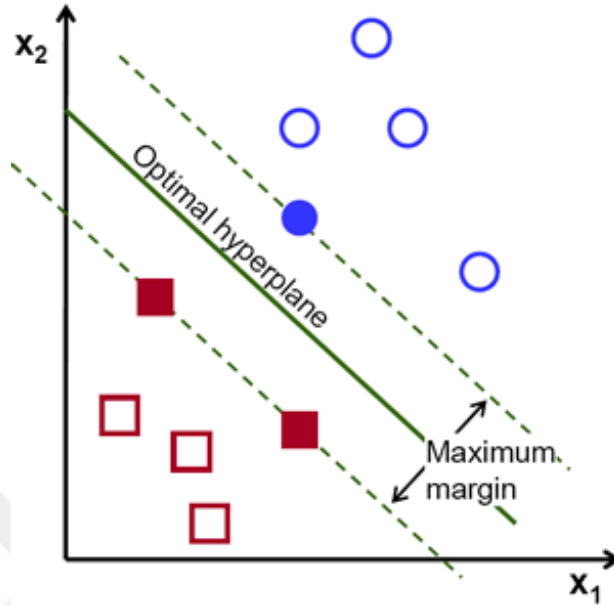


Figure 3.1: Optimal Hyperplane for SVM

The optimal hyperplane is defined by the maximum distance of the samples to the hyperplane. Hence, linear SVM becomes a maximization problem for the best separating margin. The samples that define the hyperplane and nearest to it, are named support vectors. Any possible hyperplane is defined by the equation 3.18.

$$w \cdot x + b = 0 \quad (3.18)$$

The optimal hyperplane is the minimization of equation 3.19

$$\Phi = w \cdot w \quad (3.19)$$

subject to the constraints in equation 3.20

$$y_i(x_i \cdot w + b) \geq 1, i = 1, 2, \dots, l. \quad (3.20)$$

where y_i is the output and x_i is the input of the given dataset, and is completed by using Lagrangian methods. The Lagrangian defined for the optimization is

$$L(w, b, \Lambda) = \frac{1}{2} w \cdot w \sum_{i=1}^l \alpha_i [y_i (x_i \cdot w + b) - 1] \quad (3.21)$$

where $\Lambda^T = (\alpha_1, \dots, \alpha_l)$. The objective of Lagrangian optimization is maximizing the w and b while minimizing α_i 's. As a result of this optimization, optimal hyperplane can be defined by equation 3.22

$$w = \sum_{i=1}^l \alpha_i y_i x_i + b, \alpha_i > 0 \quad (3.22)$$

In the case of no optimal separating hyperplane, classifier may need to ease the constraints on the optimization problem. Therefore, ξ_i slack variables are introduced and the minimization of equation 3.19 becomes

$$\Phi = w \cdot w + C \sum_{i=1}^l \xi_i \quad (3.23)$$

subject to the constraints 3.24 and 3.25

$$\xi_i \geq 0, i = 1, 2, \dots, l. \quad (3.24)$$

and

$$y_i ((w \cdot x_i) + b) \geq 1 - \xi_i, i = 1, 2, \dots, l. \quad (3.25)$$

In the case of nonlinear and non-separable data, kernel function method is used to perform classification. Input samples are mapped to a higher dimensional feature space by using several kernel functions [69].

In this thesis, LibSVM is used for the implementation of SVM. LibSVM is a library that provides SVM for binary and multi class classification and regression on several platforms such as MATLAB, Java, and Python. There are two types of SVM for classification provided by LibSVM, C SVM and ν SVM. The difference between these SVMs is only the cost parameter. C SVM uses C as cost parameter and it is a soft-margin classifier. C is usually defined in the logarithmic scale and can be too small or too high but it cannot be smaller than 0. Defining a big value for C makes the classifier prone to errors. When the cost value is big, classifier has the luxury to define new samples to wrong classes. If the C value is small, it becomes a hard-

margin classifier and too strict. This can be a bad thing too. Therefore, an optimization for classifier parameters is a must when performing classification. ν SVM uses ν as cost parameter and solve the optimization problem of equation 3.26

$$\Phi = w \cdot w - \nu\rho + \frac{1}{m} \sum_{i=1}^m \xi_i \quad (3.26)$$

subject to the constraints

$$y_i((w \cdot x_i) + b) \geq \rho - \xi_i, i = 1, 2, \dots, l. \quad (3.27)$$

$$\xi_i \geq 0 \quad (3.28)$$

$$\rho \geq 0 \quad (3.29)$$

ν is restricted in the range of [0 1] and a hard-margin classifier [70].

Four different kernel functions are applicable in LibSVM; linear, polynomial, radial basis function (RBF) and sigmoid. Several parameter changes may affect the performance of these SVM types. Linear SVM is defined by the equation 3.30

$$K(x_i, x_j) = x_i^T x_j \quad (3.30)$$

and only the C parameter change may affect the classification performance. Polynomial SVM is defined by the equation 3.31.

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d \quad (3.31)$$

C, γ , r coefficient and degree of the polynomial function affects the classification performance. In RBF SVM, radial basis function defined by the equation 3.32.

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}, \gamma > 0 \quad (3.32)$$

is used as kernel function. C and γ parameters are important for the classification performance of RBF SVM. Lastly, sigmoid function defined by the equation 3.33.

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \quad (3.33)$$

is used as the kernel for this type of SVM. C , γ and r coefficient is used to optimize the classification performance of this type of SVM [71].

SVMs are prone to big data. Therefore, an optimal classification performance for gene expression or proteomics data can be achieved with or without any pre-processing or gene selection [72], [73].

3.2.2 Multi-Layer Perceptrons

A multilayer perceptron is a type of artificial neural network (ANN). The basic principle of artificial neural networks is to estimate the outputs by using inputs to a set of interconnected nodes and adjusting the parameters of these nodes to minimize the errors in output. The simplest form of ANN includes an output layer. Number of nodes in this layer is related with the number of input samples and the number of output classes. This type of ANN is named single-layer perceptron or linear perceptron and it is not applicable to nonlinear data. Therefore, one or more hidden layers are introduced to perform classification for nonlinear data as well. When an ANN includes hidden layers apart from output layer, it is called multi-layer. The number of layers and the neurons in these layers are not specified and may change depending on the classification problem.

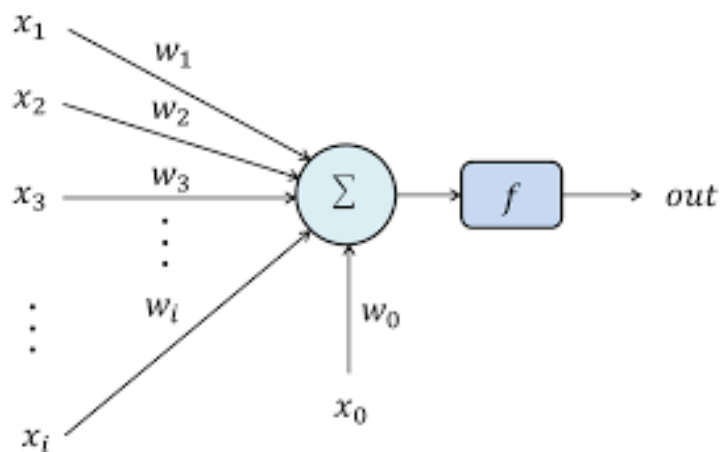


Figure 3.2 Linear Perceptron

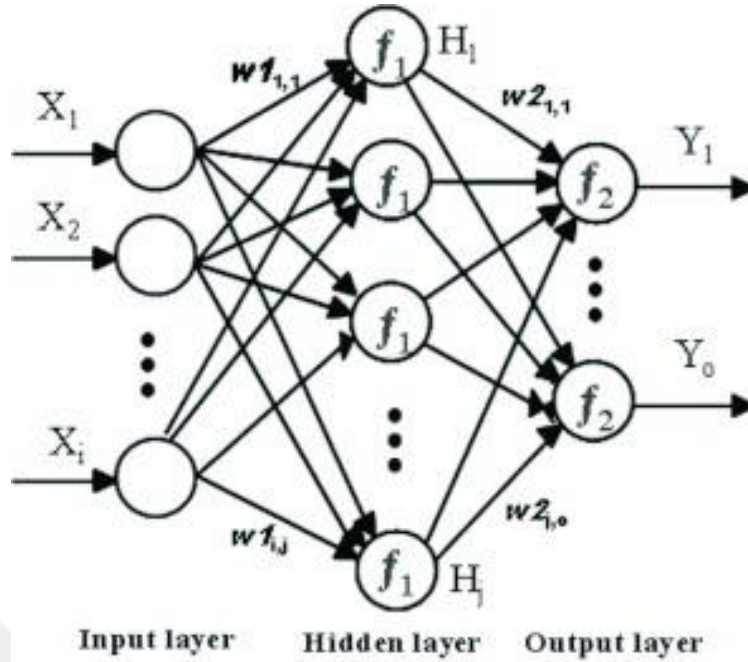


Figure 3.3: Multi-layer Perceptron

Linear perceptrons input i samples and a bias term (b or x_0), assign them weights and then output the result by summing each input and its weight through an activation function. In the case of MLP, each layer's output becomes the input for the next layer and perceptron learns by updating the parameters of layers using a learning algorithm. Back-propagation learning algorithm is used in multilayer perceptrons [74], [75].

Back-propagation learning algorithm initializes the parameters of the nodes, weights and bias, updates these parameters to minimize the output error. It can be applied with two different approaches; batch and online mode. In the batch mode, whole sample set is used and average values are aggregated to update the parameters. In online mode, the parameters are updated for every sample. Therefore, the most important difference between these modes is the computation time. If the data is too big, using all samples to update the parameters will take time. Further, Online learning is better and more reliable than the average values of batch mode learning [76]. An error function is used to minimize the output error which is usually least mean square error function. MLP is defined by the equation 3.34.

$$y = \varphi\left(\sum_{i=1}^n w_i x_i + b\right) = \varphi(w^T x + b) \quad (3.34)$$

where w is the weight matrix, x is the input samples, b is the bias term, y is the output and δ is the activation function. The back-propagation algorithm uses the gradient descent rule (3.35) to update the node parameters and minimize the output error.

$$w_{ij}(n+1) = w_{ij}(n) - \eta \frac{\partial E}{\partial w_{ij}} \quad (3.35)$$

where w is the weight, η is the learning rate and E is the error function in equation 3.36 and 3.37.

$$E = \frac{1}{2} \sum_{i=1}^n E(n) \quad (3.36)$$

$$E(n) = \|y(n) - y^L\|^2 \quad (3.37)$$

where y^L is the output of MLP when the input is $x(n)$.

The activation function may change according to the classification problem. In the case of linear classification problems, a linear activation function is used and input-output relations are defined by linear algebraic equations. Generally, two types of activation functions are used: hyperbolic tangent sigmoid (3.38) and logistic sigmoid function (3.39).

$$y(x) = \tanh(x) \quad (3.38)$$

$$y(x) = (1 + e^{-x})^{-1} \quad (3.39)$$

Radial basis functions can be used as activation functions as well.

In this thesis, a simple feedforward back-propagation perceptron is employed to perform classification for different types of cancer. This multilayer perceptron with three layers is set up in MATLAB. The first hidden layer included 16 neurons and the second one included 8 neurons. The number of neurons is determined by trial and error. Levenberg-Marquadt as training function and mean squared error as performance function are chosen as the characteristic functions of perceptron. Dimensionality is a big problem for multilayer perceptrons and, they are not applicable to big data. Classification of the microarray datasets is feasible only after applying feature selection and reducing the size of the data.

Chapter 4

Experimental Results

Experimental results for gene selection and classification of three different datasets is completed and reviewed in this chapter. Gene selection is performed with nine different feature selection algorithms. Two classification algorithms are used to classify and evaluate the suitability of the selected gene subsets. Microarray datasets provide a great quantity of data about genes, but their sample and attribute sizes are too large to analyze before any type of pre-processing. Even though some types of classifiers are applicable for this kind of big data, we still have to define appropriate gene subset for biological relevancy part of the study.

Three different publicly available microarray datasets used in this study to classify cancerous tissues and cancer types. All three datasets consist of binary classification problems. First dataset is Golub's leukemia microarray dataset which contains 7,129 gene expression values from 72 patients as features from two different leukemia types, Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). High density oligonucleotide microarray of Affymetrix is used to produce gene chip by using the bone marrow cell samples [40]. Second dataset is prostate cancer microarray dataset which contains 12,600 gene expression values as features derived from 102 patients who have normal or cancerous prostate tissue samples. Affymetrix GeneChip is used to produce the microarray for prostate dataset as well [45]. Third dataset is diffuse large B-cell lymphoma (DLBCL) microarray dataset which contains 7,129 gene expression values as features from 77 patients to classify diffuse large B-cell lymphoma and Follicular Lymphoma (FL). Gene

expression values of DLBCL dataset is measured by using Affymetrix GeneChip [47].

Table 4.1: Datasets' Description

Reference Study	Number of samples (Total)	Number of Genes	Number of Samples (per class)		Production Platform
(T.R. Golub, 1999)	72	7,129	47 ALL	25 AML	Affymetrix
(Dinest Singh, March 2002)	102	12,600	52 Tumor	50 Normal	Affymetrix
(Margaret A. Shipp, January 2002)	77	7,129	58 DLBCL	19 FL	Affymetrix

SVM provided by LibSVM library and multi-layer perceptron are used to evaluate the performance of feature selection algorithms. All of the experiments conducted on MATLAB. 10-fold cross validation is used to separate the data as train and test sets. Some of the feature selection algorithms provide ranking of the genes and some of them provide fixed-number gene subsets. Since ranking feature selection algorithms required less computational time, they are evaluated for each of the fold in cross validation to consider the stability of the feature selection algorithms. The computation time for feature selection algorithms which provide a fixed-number gene subset were long. Therefore, a feature number is determined for these algorithms and tests are conducted over those defined gene subsets. Comparison of computational time is completed for each feature selection method using leukemia dataset. Same experimental procedure as feature ranking algorithms is applied to the feature selection algorithms that provide gene subset to show the computational time requirements of these algorithms. Related tables are provided in the “Experimental Results of Leukemia Data” sub section. Gene subset stability could not be measured for these feature selection algorithms.

First, experiments conducted on raw data to see the classifier performance without any pre-processing. Multi-layer perceptron was not able to classify this type of big data without any pre-processing. On the other hand, SVM is prone to big data and performed well for microarray datasets without any pre-processing. In SVM, each kernel has several parameters which may affect their classification performance. Therefore, after classifying the raw data with the default parameters, grid search for significant parameters is performed to see the effects of parameter change. Ranges for SVM parameters adjusted according to suggested values from previous studies [44], [77]. Since C SVM performed better than v SVM in general, C SVM is used for all experiments.

Table 4.2: Classification accuracy for leukemia data, No feature selection

SVM Type	Kernel	Accuracy
C SVM	Linear	97.32%
C SVM	Polynomial	97.32%
C SVM	RBF	65.41%
C SVM	Sigmoid	65.41%

For leukemia data, linear and polynomial SVM performed better than RBF or sigmoid SVM. When C SVM with linear kernel used, changes in the C parameter may affect the performance of classifier. Therefore, parameter search in the range of $[2^{-6} 2^{16}]$ is performed for C parameter. C change has a light effect on classification accuracy, decreases to %97.22 but this result is in the range of acceptable error.

In the case of C-SVM with polynomial kernel C, gamma, r coefficient and degree of the polynomial are the most important parameters. Therefore, grid search for these parameters performed in the range of $[2^{-6} 2^{16}]$ for C, $[2^{-16} 2^6]$ for gamma, $[2^{-6} 2^6]$ for r and [1 10] for degree of the polynomial kernel function. It is seen that changing C, gamma and r has no effect on classification accuracy but the polynomial kernels which has a degree greater than 5 results poorly for classification tasks of leukemia dataset.

For C SVM with RBF kernel, C and gamma parameters are important and may affect the classification performance. Therefore, grid search for parameters C

and gamma are performed in the range of $[2^{-6} 2^{16}]$ and $[2^{-16} 2^6]$, respectively. Changing the value for C and gamma in the specified ranges did not affect the performance of C-SVM with RBF kernel; it still performs poorly.

Lastly, C SVM with sigmoid kernel is sensitive for the changes in C, gamma and r. Therefore, grid search for these parameters conducted in the range of $[2^{-6} 2^{16}]$ for C, $[2^{-16} 2^6]$ for gamma and $[2^{-6} 2^6]$ for r coefficient. Sigmoid kernel performed poorly for this data and parameter changes did not affect the classification performance.

For the classification task of prostate cancer dataset, linear and polynomial SVMs performed better than others.

Table 4.3: Classification accuracy for prostate data, No feature selection

SVM Type	Kernel	Accuracy
C SVM	Linear	90.09%
C SVM	Polynomial	91.18%
C SVM	RBF	50.90%
C SVM	Sigmoid	50.90%

All parameter grid search steps done for the leukemia dataset is repeated for the prostate cancer dataset classification task in the previously defined parameter ranges. Adjusting the C parameter bigger than its default value, which is 1, affected the classification performance badly in linear SVM. Decreasing C did not have any effect. Therefore, default value for C accepted for this dataset.

In the polynomial kernel SVM, changing C, gamma and r had no effect on the classification performance but degree of the polynomial affected badly. Only first and second degree polynomial kernels were applicable for prostate cancer dataset.

C and gamma parameters are important for the RBF kernel but changing them had no effect on classification performance either.

In sigmoid SVM, changing C, gamma and r coefficient had no effect on classification performance.

For the classification task of DLBCL dataset, linear and polynomial SVMs performed better than others.

Table 4.4: Classification accuracy for DLBCL data, No feature selection

SVM Type	Kernel	Accuracy
C SVM	Linear	96.07%
C SVM	Polynomial	96.07%
C SVM	RBF	75.35%
C SVM	Sigmoid	75.35%

For DLBCL dataset, grid search for all the effective parameters done in the previously defined ranges. Linear SVM performed well with the default values for C parameter. Changing C did not have any effect on the classification performance.

Polynomial SVM performed well for DLBCL dataset but only up to three degree polynomial kernels, any degree greater than three decreased the classification performance. Changing C, gamma or r coefficient did not have any effect on the performance.

In the case of RBF and sigmoid SVMs, classification performance is poor for DLBCL dataset and changing C, gamma or r coefficient had no effect on the classification performance.

Since multi-layer perceptron could not perform without any type of feature/gene selection, feature selection methods applied in order to find an appropriate gene subset to classify. These subsets were useful to find genes appropriate and sufficient enough for diagnostic purposes. Nine different filter feature selection methods are used to find optimal gene subset for three different microarray datasets.

4.1 Experimental Results for Leukemia Data

Nine different feature selection methods applied to leukemia data to find an optimal gene subset. These gene subsets' classification performance evaluated using SVM and multi-layer perceptron. 10-fold cross validation is used to separate the data into train and test sets. C SVM with all possible kernels is applied to determine the classification performance. In the case of multi-layer perceptron, a four layered feed forward back-propagation perceptron with Levenberg-Marquadt as training and mean squared error as performance function is used to evaluate the classification performance. Feature number is selected regarding the previous studies and adjusted in the range of [1 100] with 9 feature increments. First, the classification performed with the raw data. Then, normalization and scaling applied to see its effect on gene selection and classification performance. Samples are normalized so that they have 0 mean, 1 standard deviation and scaled in the range of [-1 1].

Table 4.5: Computational time of T test as feature selection algorithm

Feature Selection Method	Classifier	Elapsed Time (s)
T test	Linear SVM	6,46
T test	Polynomial SVM	31,96
T test	RBF SVM	2,19
T test	Sigmoid SVM	2,13
T test	MLP	152,98

Table 4.6: T test statistic feature selection for leukemia dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-10	Linear	100%
1-82	Polynomial	98.57%
1-28	RBF	65.47%
1-10	Sigmoid	65.53%

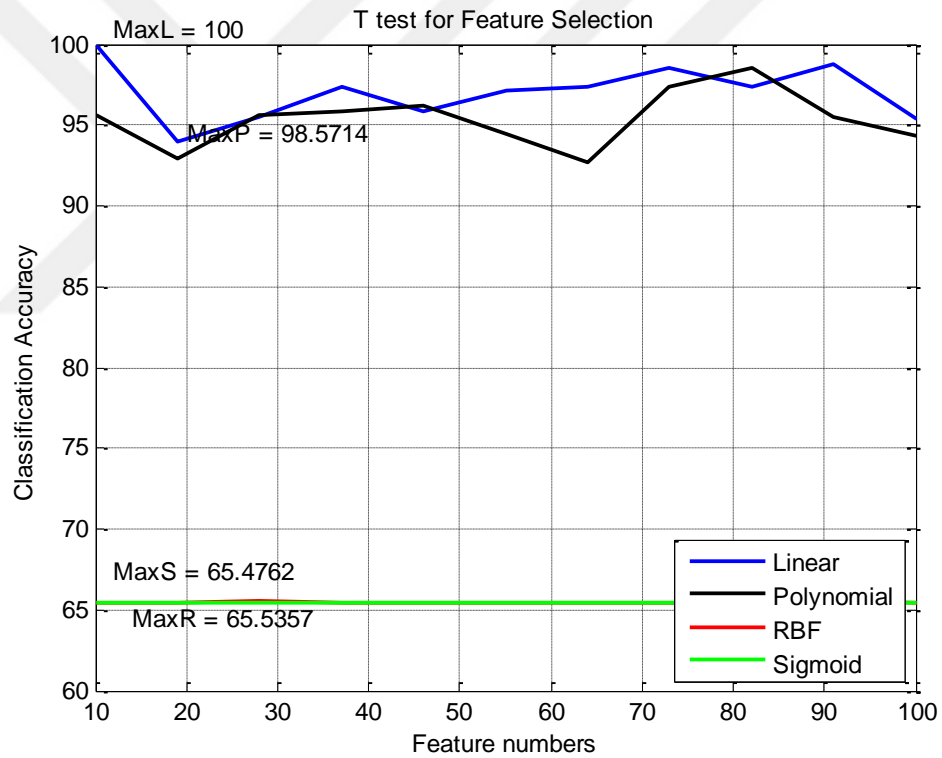


Figure 4.1: Classification performance of SVM for leukemia data (raw)

Table 4.7: T test statistic feature selection for leukemia dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-91	Linear	97.5%
1-73	Polynomial	65.53%
1-64	RBF	98.75%
1-10	Sigmoid	86.96%

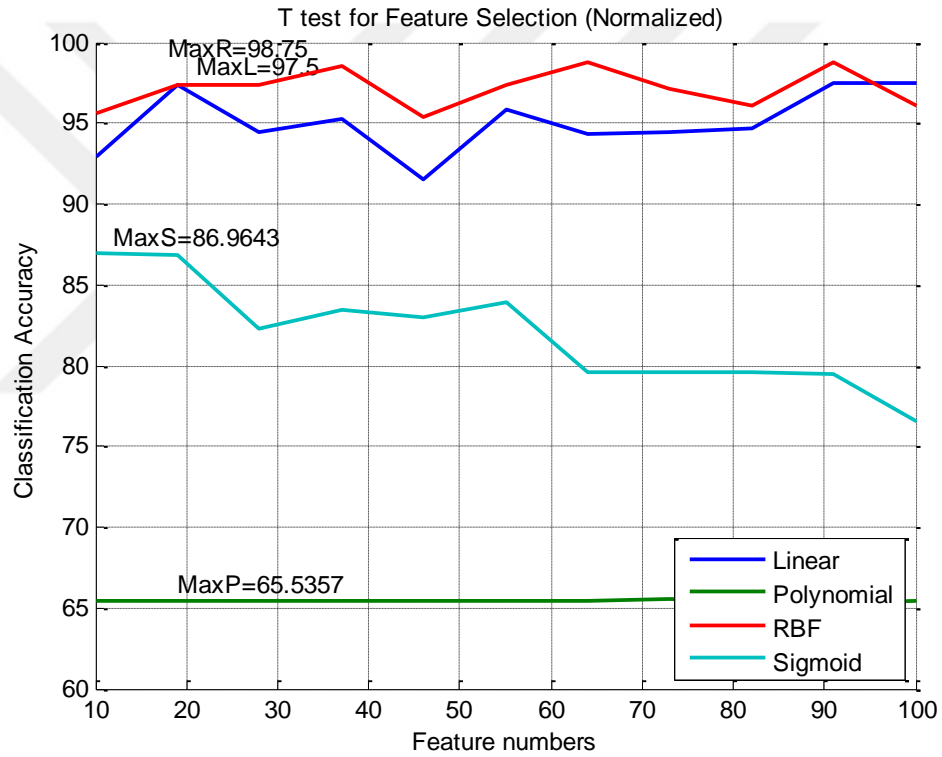


Figure 4.2: Classification performance of SVM for leukemia data (normalized)

Table 4.8: Computational time of CFS as feature selection algorithm

Feature Selection Method	Classifier	Elapsed Time (s)
CFS	Linear SVM	207,15
CFS	Polynomial SVM	227,71
CFS	RBF SVM	180,73
CFS	Sigmoid SVM	178,37
CFS	MLP	254,58

Table 4.9: CFS for leukemia dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-19	Linear	97.32%
1-55	Polynomial	96.25%
1-46	RBF	65.53%
1	Sigmoid	65.47%

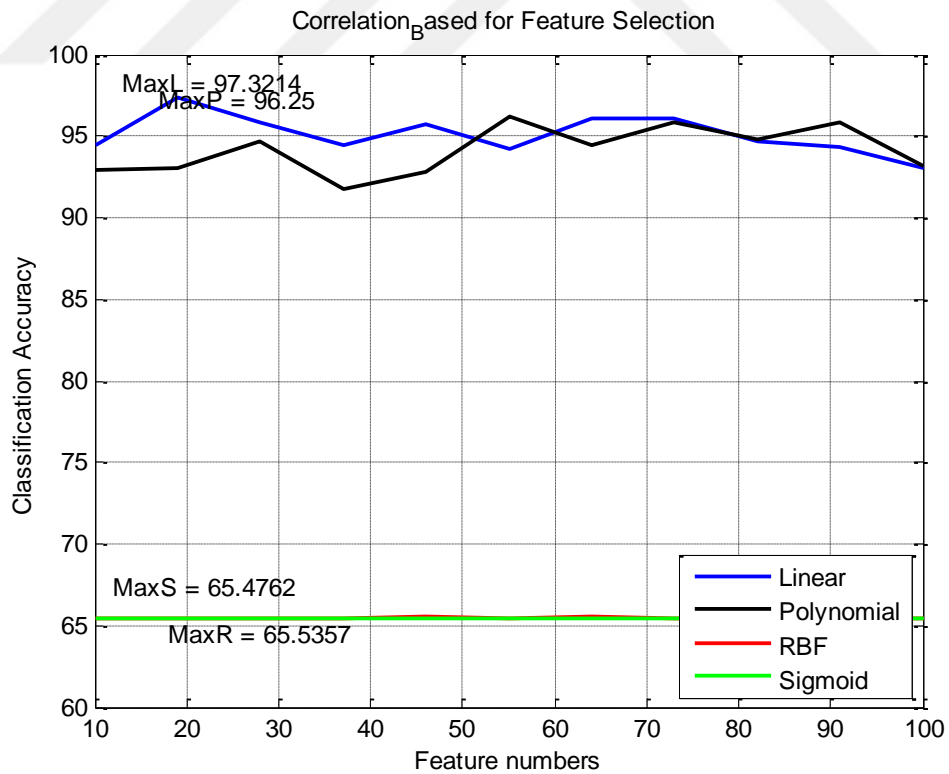


Figure 4.3: Classification performance of SVM for leukemia data (raw)

Table 4.10: CFS for leukemia dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-37	Linear	92.79%
1-19	Polynomial	65.53%
1-10	RBF	97.14%
1-19	Sigmoid	74.22%

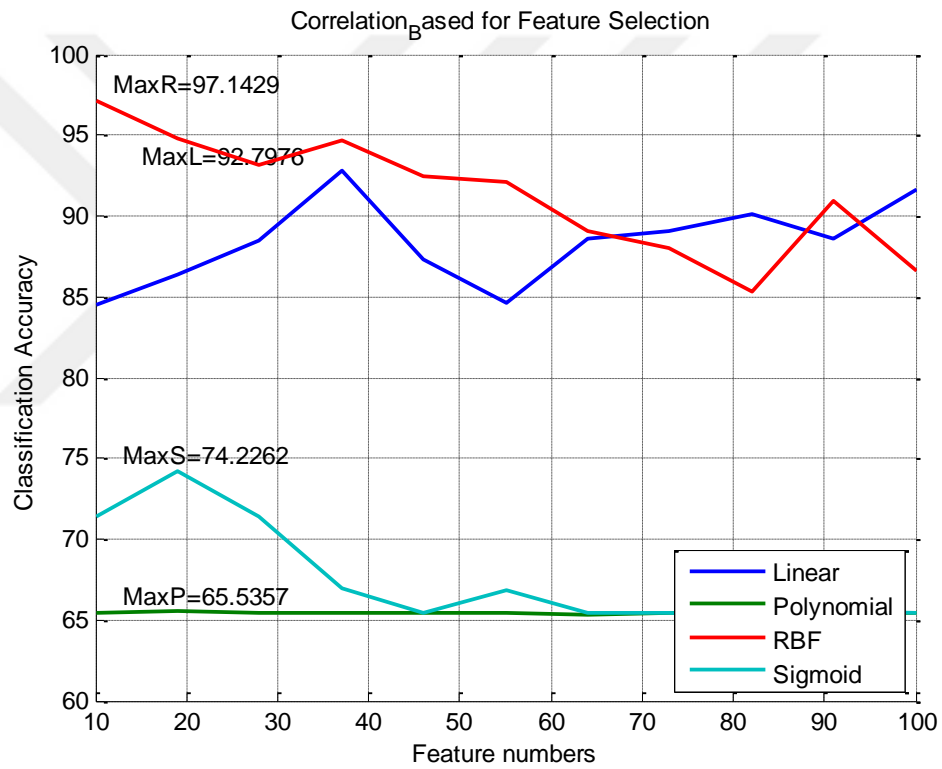


Figure 4.4: Classification performance of SVM for leukemia data (normalized)

Table 4.11: Computational time of Bhattacharyya Distance as feature selection algorithm

Feature Selection Method	Classifier	Elapsed Time (s)
Bhattacharyya Distance	Linear SVM	7,53
Bhattacharyya Distance	Polynomial SVM	39,85
Bhattacharyya Distance	RBF SVM	2,30
Bhattacharyya Distance	Sigmoid SVM	2,31
Bhattacharyya Distance	MLP	141,07

Table 4.12: Bhattacharyya Distance for leukemia dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-91	Linear	97.32%
1-73	Polynomial	88.39%
1-46	RBF	65.53%
1-28	Sigmoid	65.53%

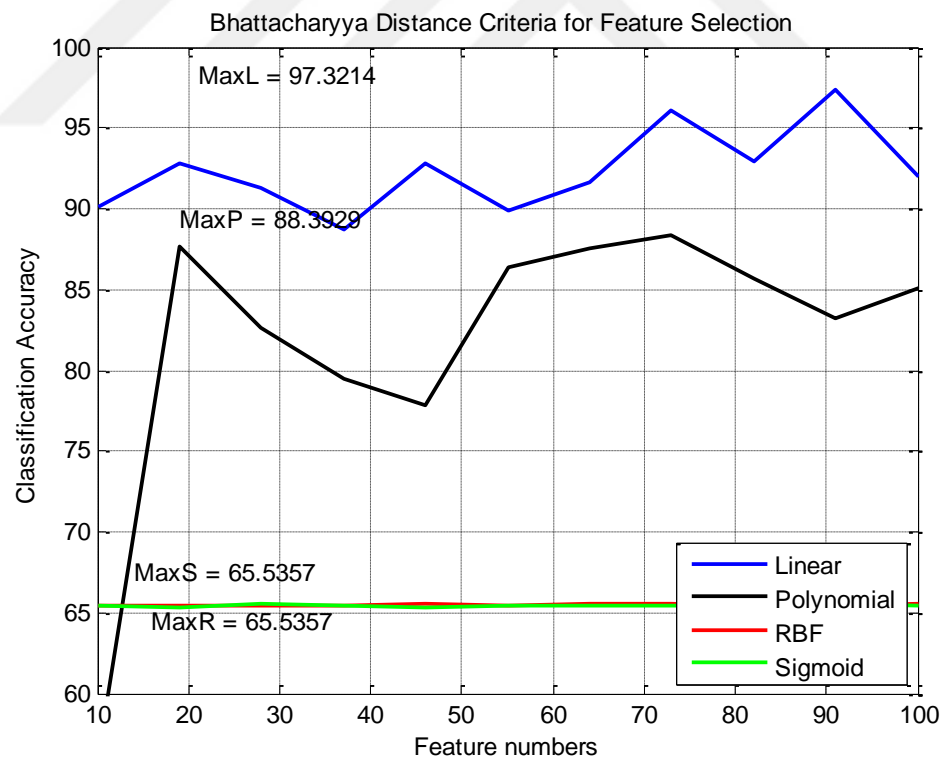


Figure 4.5: Classification performance of SVM for leukemia data (raw)

Table 4.13: Bhattacharyya Distance for leukemia dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-82	Linear	98.75%
1-10	Polynomial	65.47%
1-19	RBF	84.76%
1-10	Sigmoid	70.29%

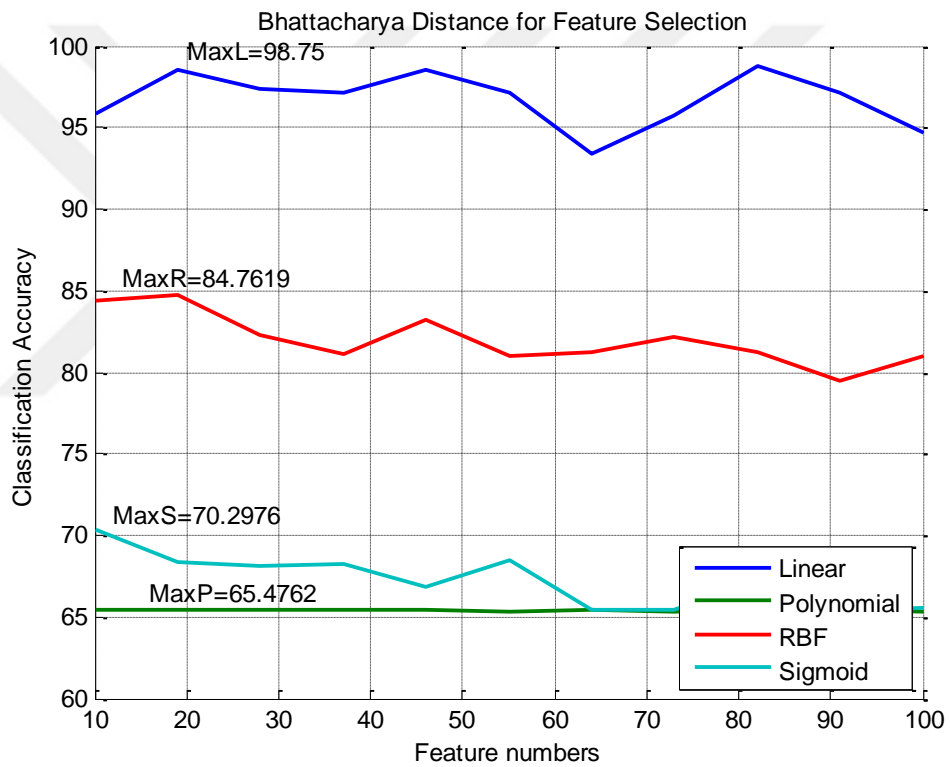


Figure 4.6: Classification performance of SVM for leukemia data (normalized)

Table 4.14: Computational time of Entropy as feature selection algorithm

Feature Selection Method	Classifier	Elapsed Time (s)
Entropy	Linear SVM	2,27
Entropy	Polynomial SVM	47,05
Entropy	RBF SVM	2,07
Entropy	Sigmoid SVM	2,10
Entropy	MLP	204,76

Table 4.15: Entropy for leukemia dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-28	Linear	100%
1-19	Polynomial	95.65%
1	RBF	66.90%
1-10	Sigmoid	77.5%

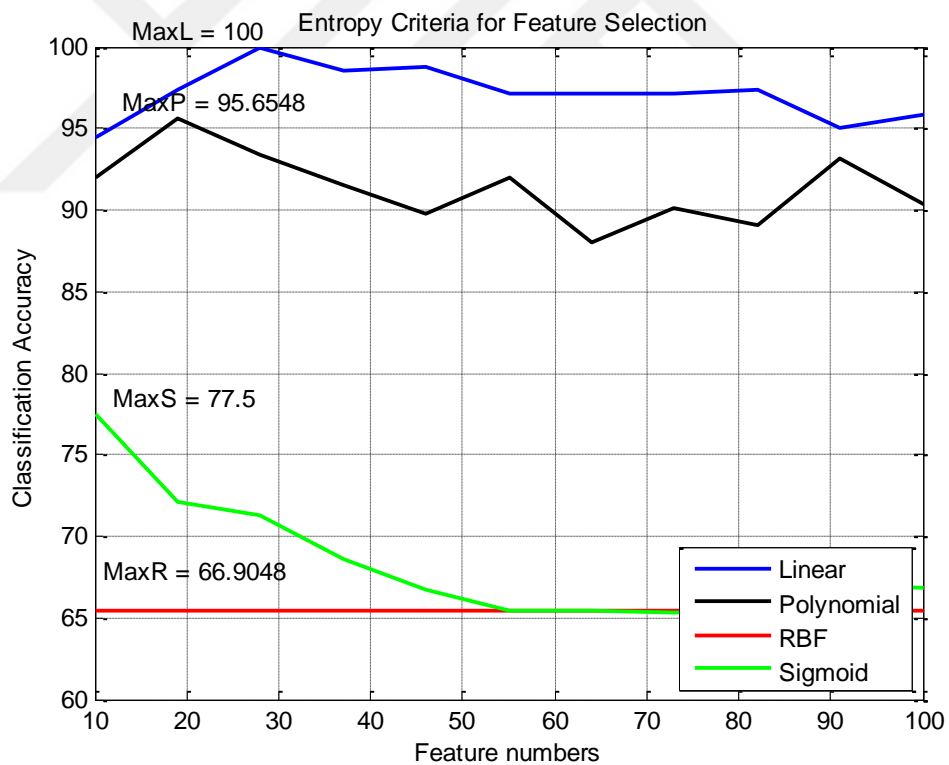


Figure 4.7: Classification performance of SVM for leukemia data (raw)

Table 4.16: Entropy for leukemia dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-28	Linear	97.5%
1	Polynomial	65.53%
1-19	RBF	87.85%
1-10	Sigmoid	72.67%

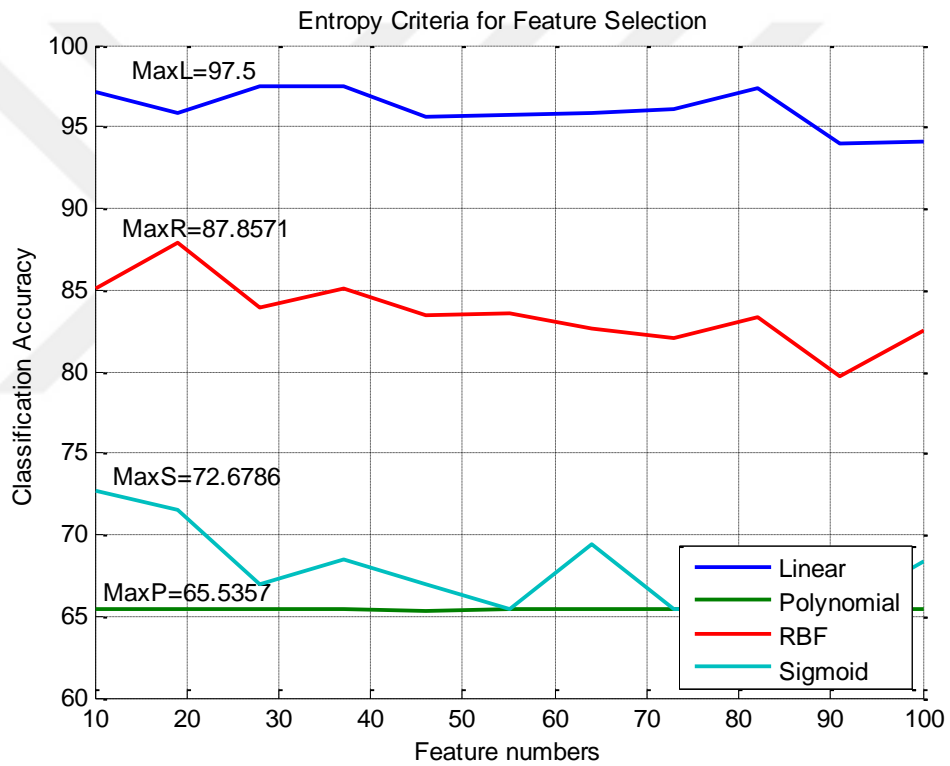


Figure 4.8: Classification performance of SVM for leukemia data (normalized)

Table 4.17: Computational time of ReliefF as feature selection algorithm

Feature Selection Method	Classifier	Elapsed Time (s)
ReliefF	Linear SVM	499,89
ReliefF	Polynomial SVM	536,36
ReliefF	RBF SVM	516,46
ReliefF	Sigmoid SVM	497,47
ReliefF	MLP	613,25

Table 4.18: ReliefF for leukemia dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	97.14%
1-46	Polynomial	95.89%
1-10	RBF	65.53%
1	Sigmoid	65.47%

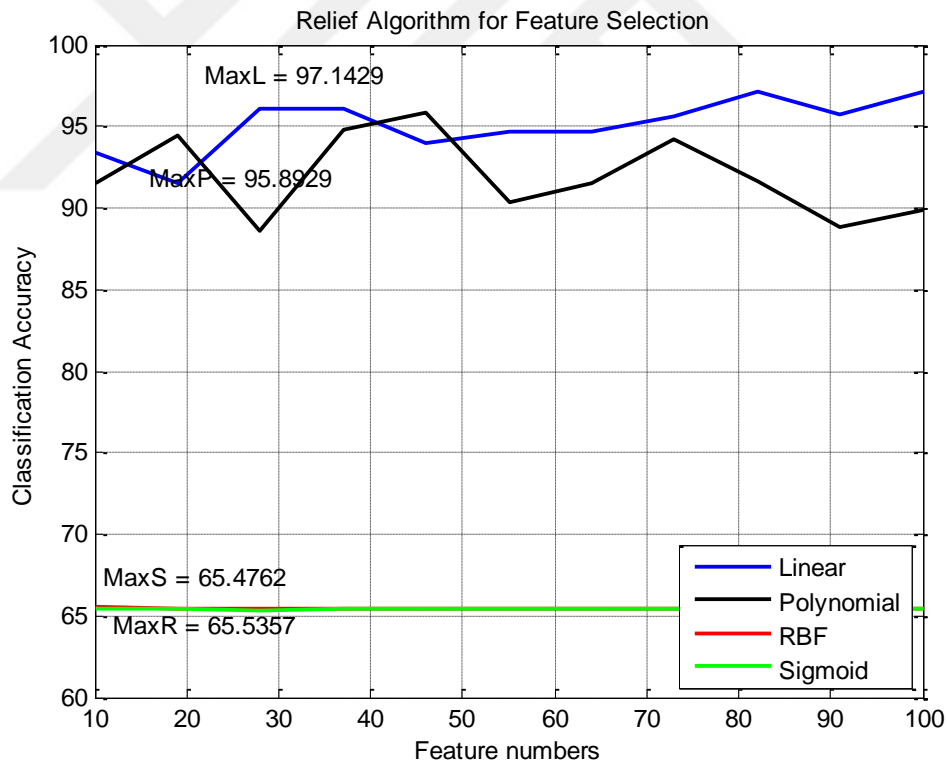


Figure 4.9: Classification performance of SVM for leukemia data (raw)

Table 4.19: ReliefF for leukemia dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-10	Linear	97.32%
1	Polynomial	65.47%
1-19	RBF	94.64%
1-10	Sigmoid	89.04%

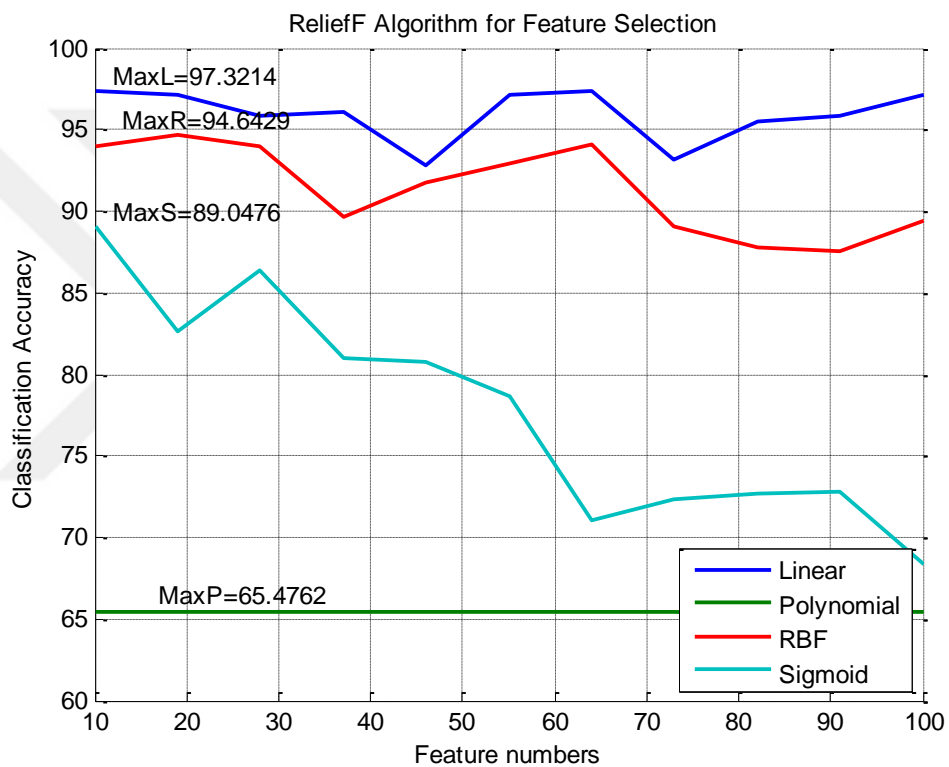


Figure 4.10: Classification performance of SVM for leukemia data (normalized)

Table 4.20: Computational time of Wilcoxon signed-rank test as feature selection algorithm

Feature Selection Method	Classifier	Elapsed Time (s)
Wilcoxon	Linear SVM	289,47
Wilcoxon	Polynomial SVM	314,77
Wilcoxon	RBF SVM	285,10
Wilcoxon	Sigmoid SVM	285,10
Wilcoxon	MLP	360,93

Table 4.21: Wilcoxon signed-rank test for leukemia dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-46	Linear	98.75%
1-28	Polynomial	94.22%
1-10	RBF	65.53%
1-28	Sigmoid	65.53%

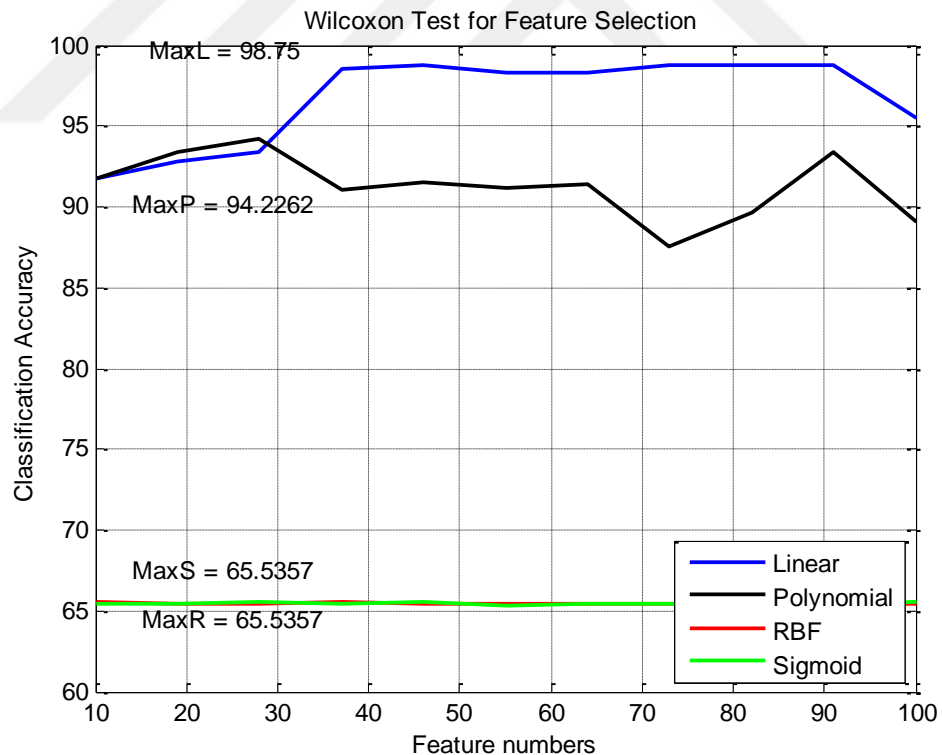


Figure 4.11: Classification performance of SVM for leukemia data (raw)

Table 4.22: Wilcoxon signed-rank test for leukemia dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-55	Linear	98.75%
1-64	Polynomial	65.53%
1-82	RBF	97.5%
1-46	Sigmoid	93.21%

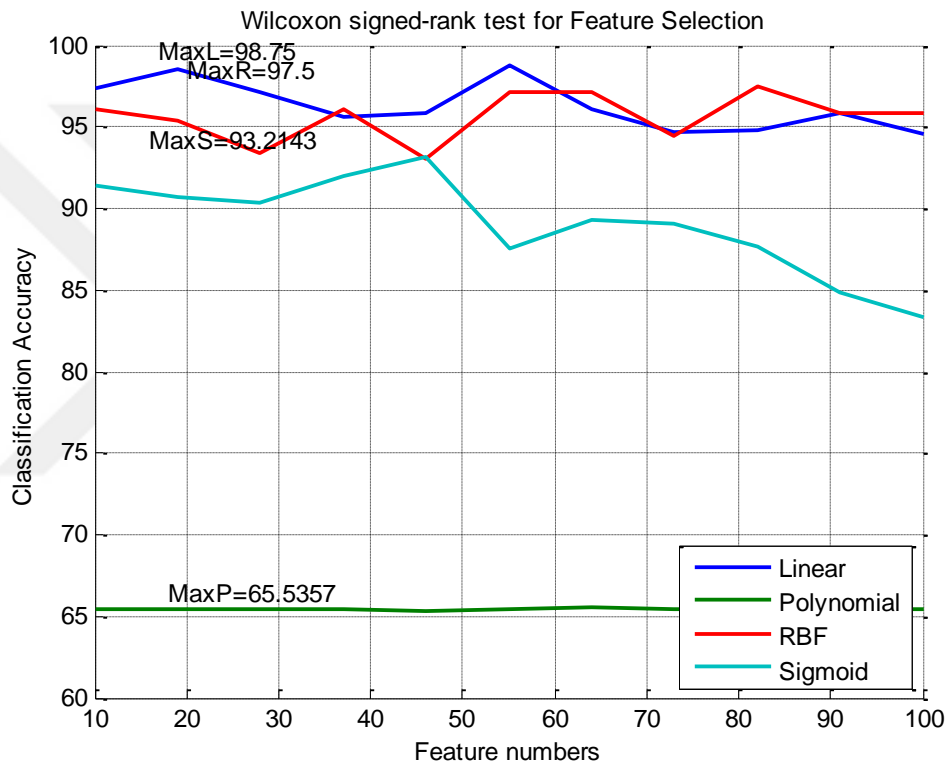


Figure 4.12: Classification performance of SVM for leukemia data (normalized)

Table 4.23: Computational time of mRMR as feature selection algorithm

Feature Selection Method	Classifier	Elapsed Time (s)
mRMR	Linear SVM	43579,75
mRMR	Polynomial SVM	44310,25
mRMR	RBF SVM	45533,68
mRMR	Sigmoid SVM	42550,60
mRMR	MLP	47745,76

Table 4.24: mRMR for leukemia dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-64	Linear	80.47%
1-37	Polynomial	77.26%
1	RBF	67.97%
1-37	Sigmoid	66.90%

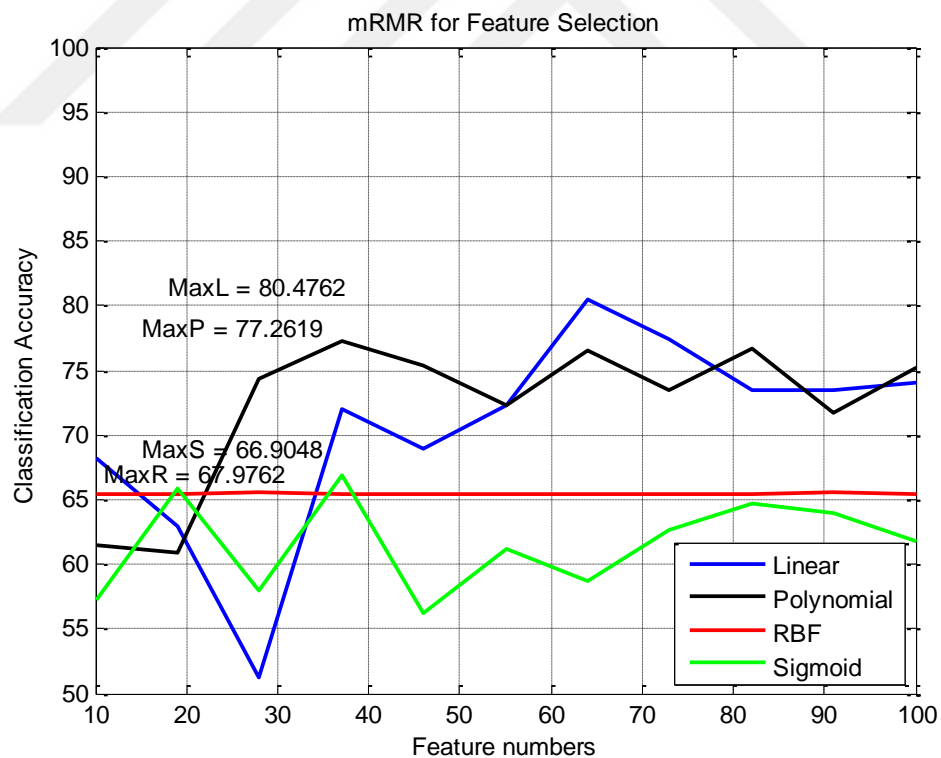


Figure 4.13: Classification performance of SVM for leukemia data (raw)

Table 4.25: mRMR for leukemia dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-73	Linear	88.03%
1-46	Polynomial	65.53%
1-19	RBF	65.53%
1-10	Sigmoid	65.47%

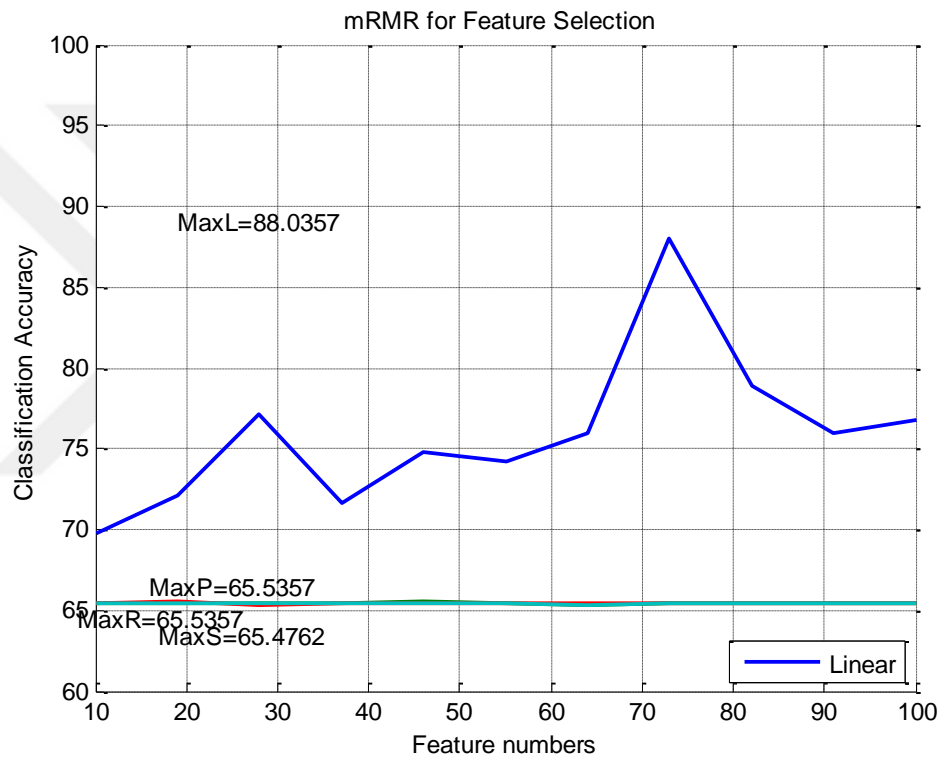


Figure 4.14: Classification performance of SVM for leukemia data (normalized)

Table 4.26: Computational time of DISR as feature selection algorithm

Feature Selection Method	Classifier	Elapsed Time (s)
DISR	Linear SVM	4225,40
DISR	Polynomial SVM	4341,12
DISR	RBF SVM	4420,40
DISR	Sigmoid SVM	4320,16
DISR	MLP	3774,57

Table 4.27: DISR for leukemia dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	85.83%
1-100	Polynomial	87.97%
1	RBF	69.34%
1-91	Sigmoid	65.53%

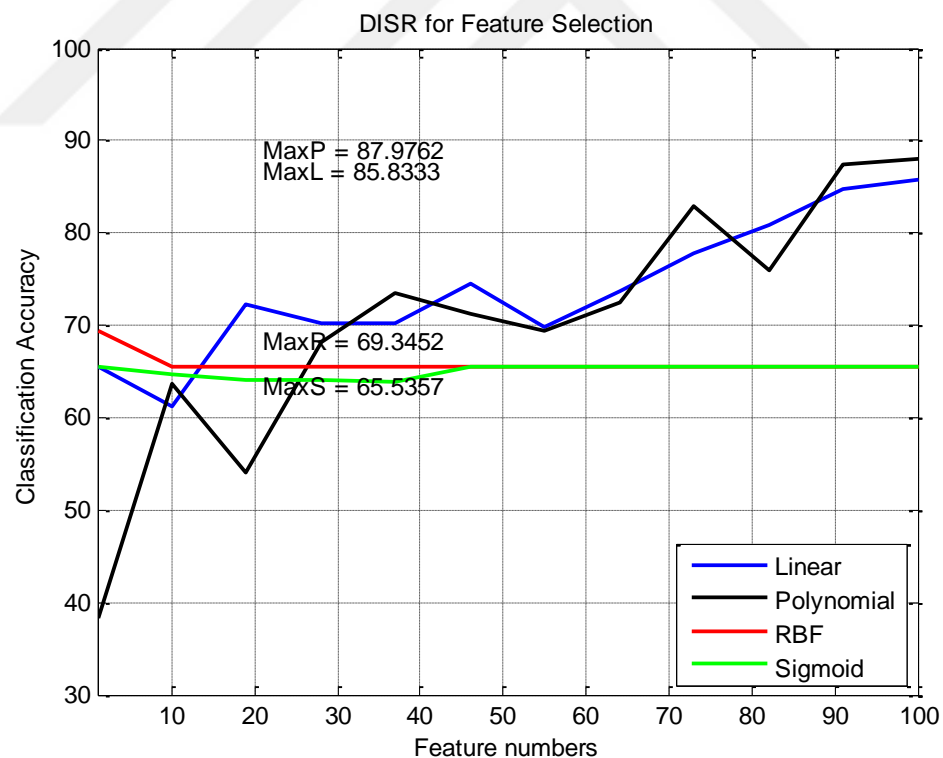


Figure 4.15: Classification performance of SVM for leukemia data (raw)

Table 4.28: DISR for leukemia dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	86.07%
1-19	Polynomial	65.53%
1-28	RBF	65.47%
1-46	Sigmoid	65.53%

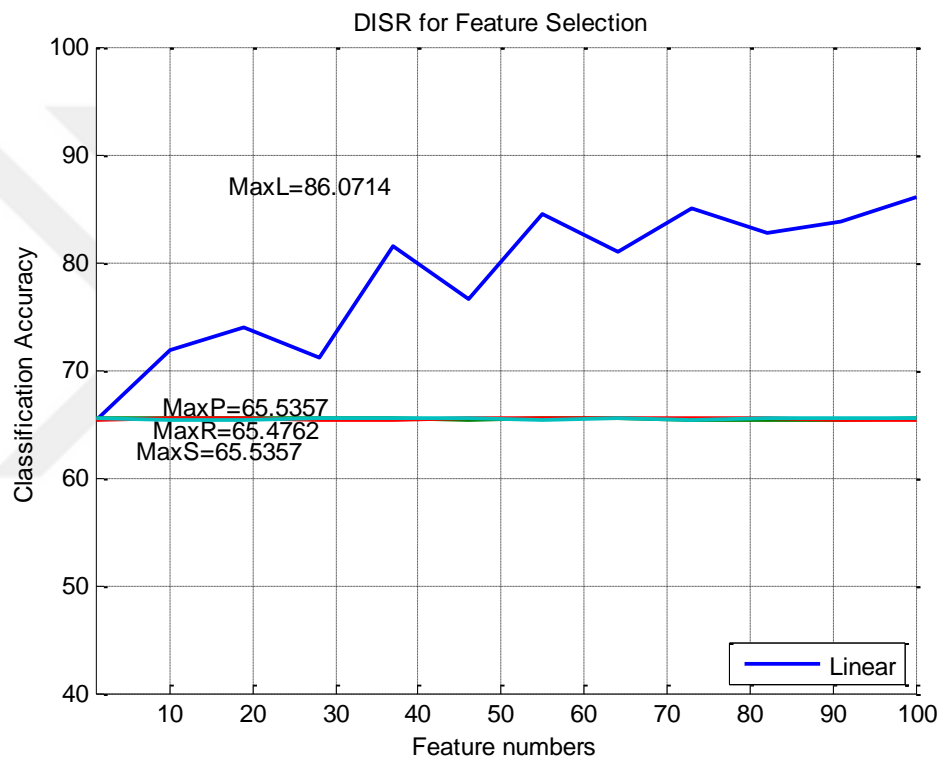


Figure 4.16: Classification performance of SVM for leukemia data (normalized)

Table 4.29: Computational time of ROC as feature selection algorithm

Feature Selection Method	Classifier	Elapsed Time (s)
ROC	Linear SVM	5,56
ROC	Polynomial SVM	31,92
ROC	RBF SVM	5,34
ROC	Sigmoid SVM	5,27
ROC	MLP	207,39

Table 4.30: ROC for leukemia dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-28	Linear	98.75%
1-82	Polynomial	98.75%
1-28	RBF	65.53%
1-10	Sigmoid	65.47%

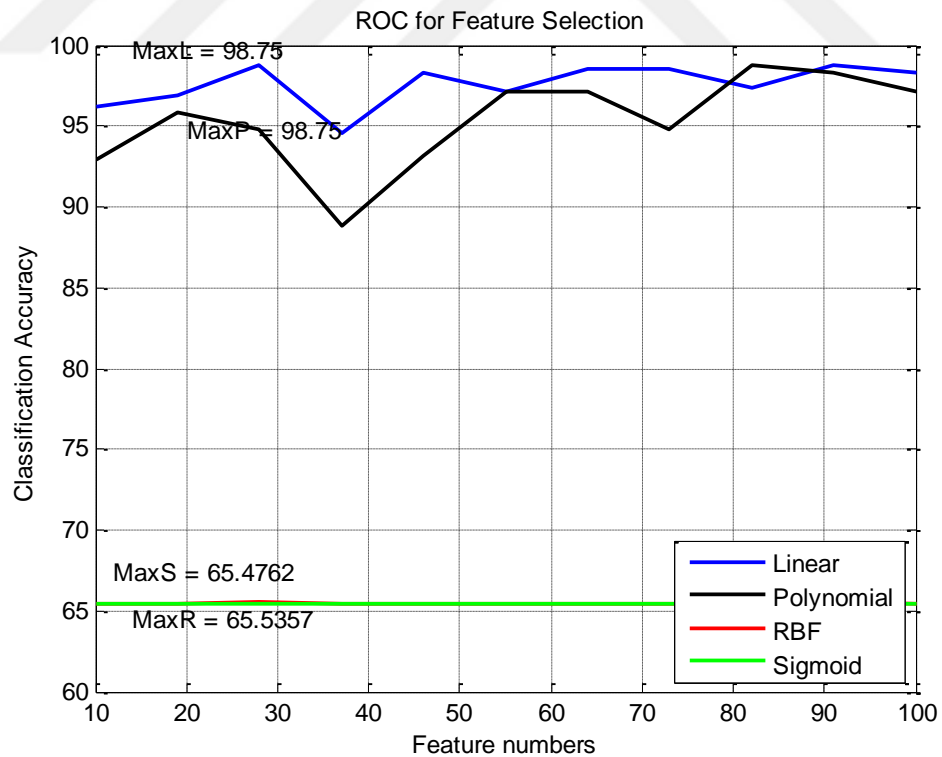


Figure 4.17: Classification performance of SVM for leukemia data (raw)

Table 4.31: ROC for leukemia dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-19	Linear	98.75%
1	Polynomial	65.53%
1-28	RBF	98.75%
1-10	Sigmoid	92.97%

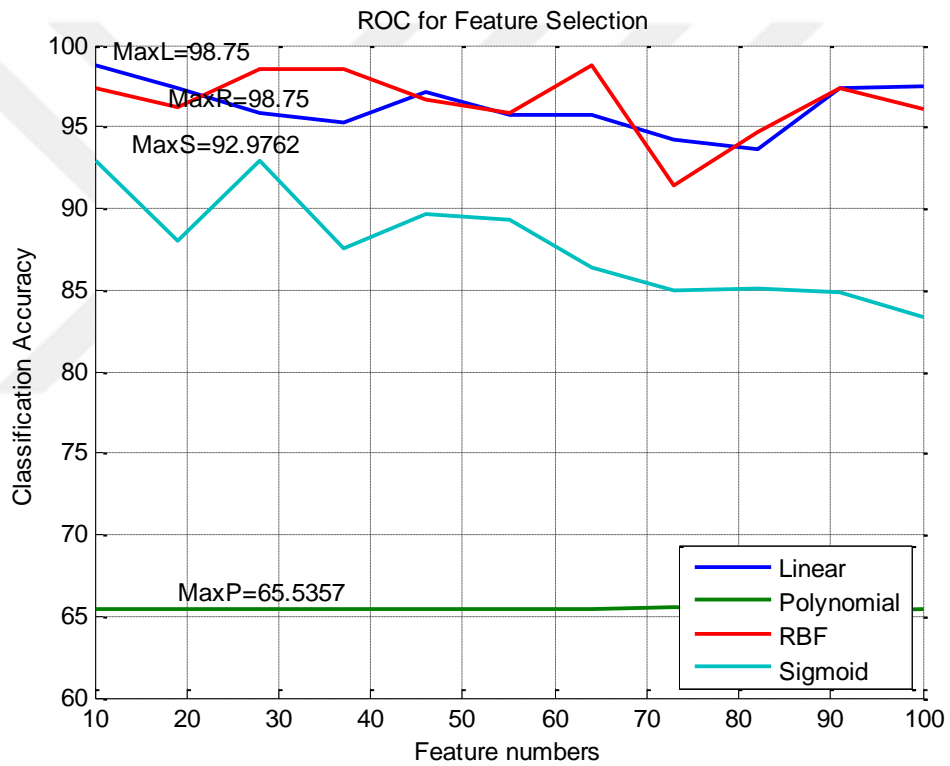


Figure 4.18: Classification performance of SVM for leukemia data (normalized)

Table 4.32: Feature numbers for leukemia dataset using MLP (raw)

Number of features	Feature Selection Algorithm	Classification Accuracy
1-80	T test	97.5%
1-80	ROC	96.90%
1-30	ReliefF	91.60%
1-50	Wilcoxon	95.83%
1-10	CFS	100%
1-10	Entropy	97.14%
1-90	MRMR	94.04%
1-90	DISR	94.46%
1-50	Bhattacharyya	95.83%

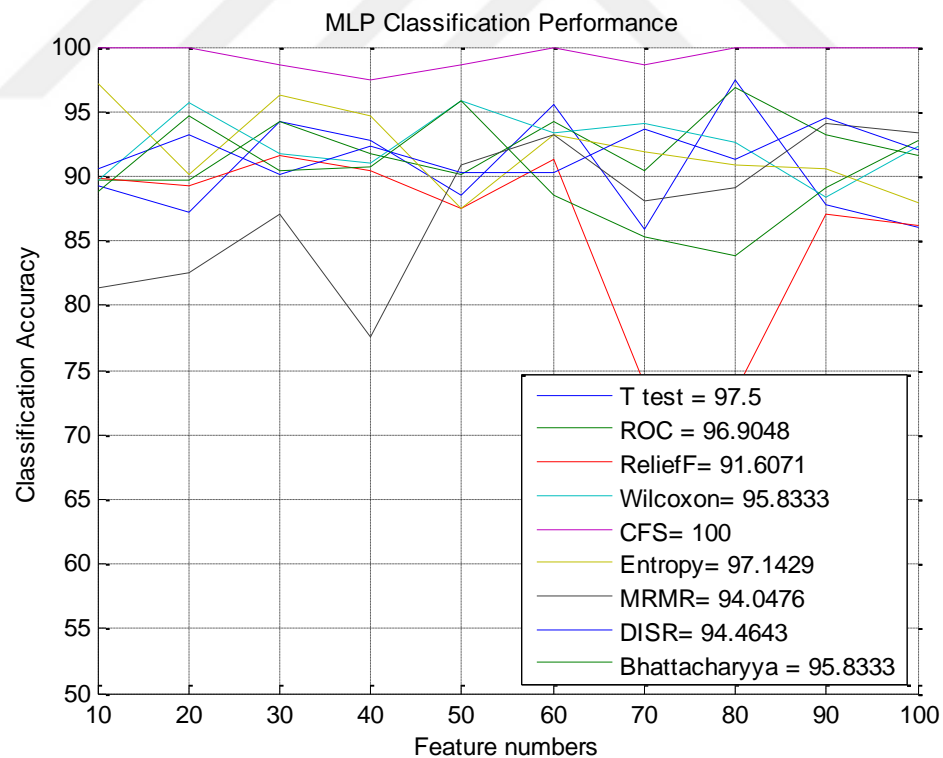


Figure 4.19: Classification performance of MLP for leukemia data (raw)

Table 4.33: Feature numbers for leukemia dataset using MLP (normalized)

Number of features	Feature Selection Algorithm	Classification Accuracy
1-80	T test	94.46%
1-20	ROC	95.47%
1-60	ReliefF	91.90%
1-100	Wilcoxon	94.28%
1-40	CFS	96.90%
1-100	Entropy	95.65%
1-70	MRMR	82.67%
1-90	DISR	90.71%
1-100	Bhattacharyya	93.15%

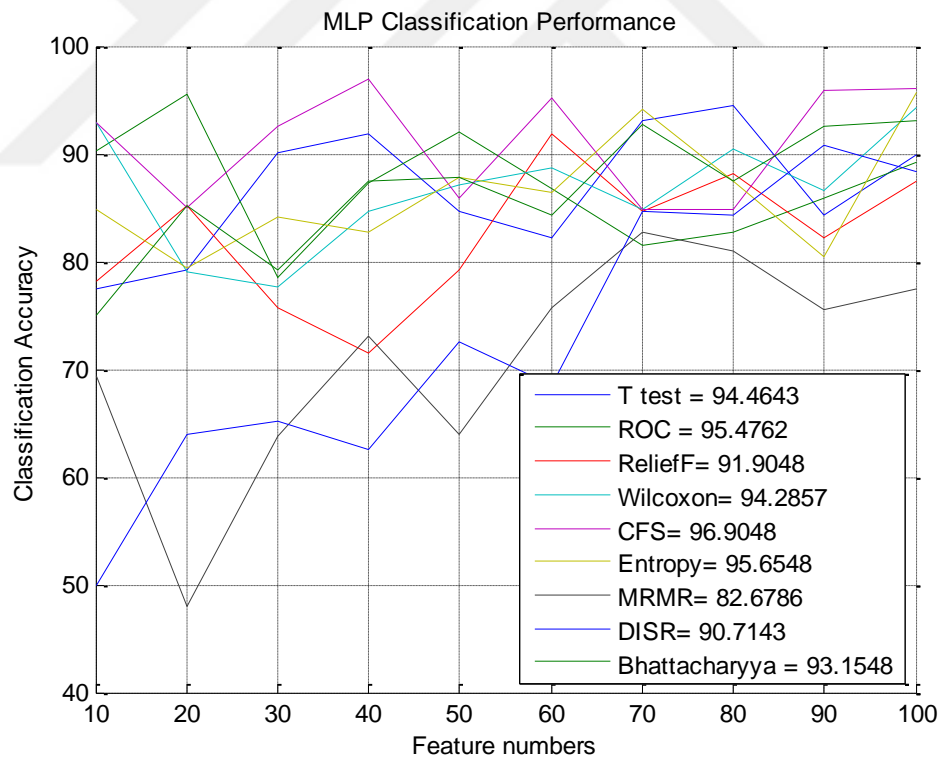


Figure 4.20: Classification performance of MLP for leukemia data (normalized)

4.2 Experimental Results for Prostate Cancer Data

Same experimental procedure followed for the prostate data. Nine different feature selection methods applied to find an optimal gene subset. Classification performance for the gene subsets evaluated using SVM and multi-layer perceptron. 10-fold cross validation is used to separate the data into train and test sets. In the case of SVM, C SVM with four different kernels is applied to determine the classification performance. A four layered feed forward back-propagation perceptron with Levenberg-Marquadt as training and mean squared error as performance function is used to evaluate the classification performance. Feature number for the gene subset is selected according to previous studies and adjusted in the range of [1 100] and experiments conducted with 9 feature increments. First, the classification performed with the raw data. Then, normalization and scaling so that the data will have 0 mean, 1 standard deviation and scaled in the range of [-1 1], applied to see its effect on gene selection and classification performance.

Table 4.34: T test statistic feature selection for prostate dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	92.09%
1-28	Polynomial	92.18%
1	RBF	78.36%
1-10	Sigmoid	51.90%

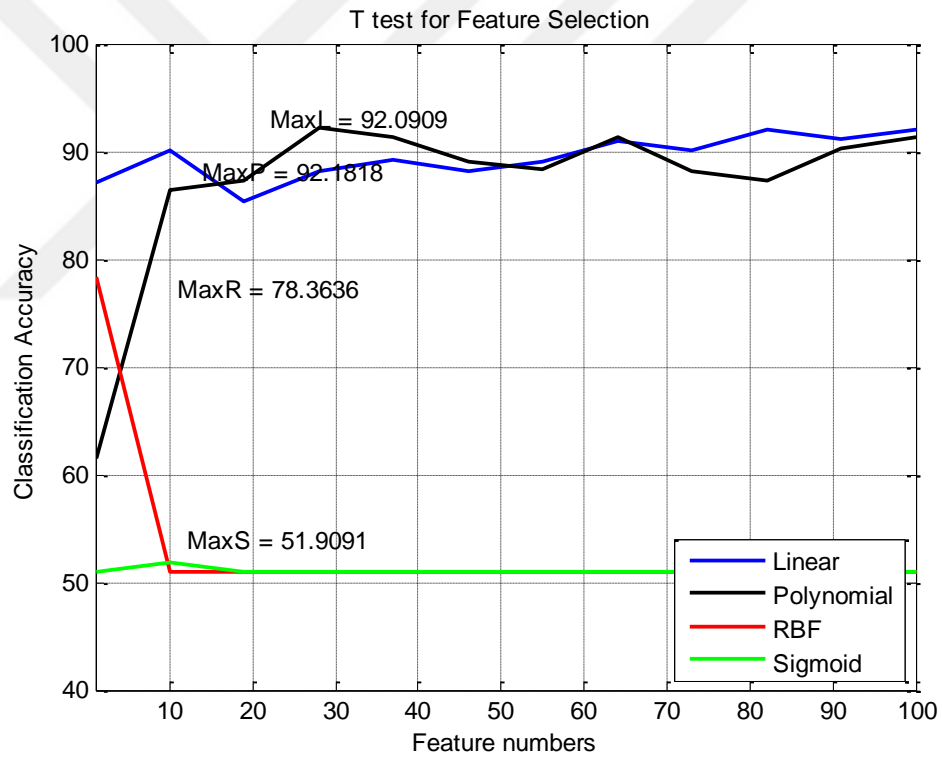


Figure 4.21: Classification performance of SVM for prostate data (raw)

Table 4.35: T test statistic feature selection for prostate dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-82	Linear	95.18%
1	Polynomial	50.90%
1-19	RBF	92.36%
1-10	Sigmoid	92.18%

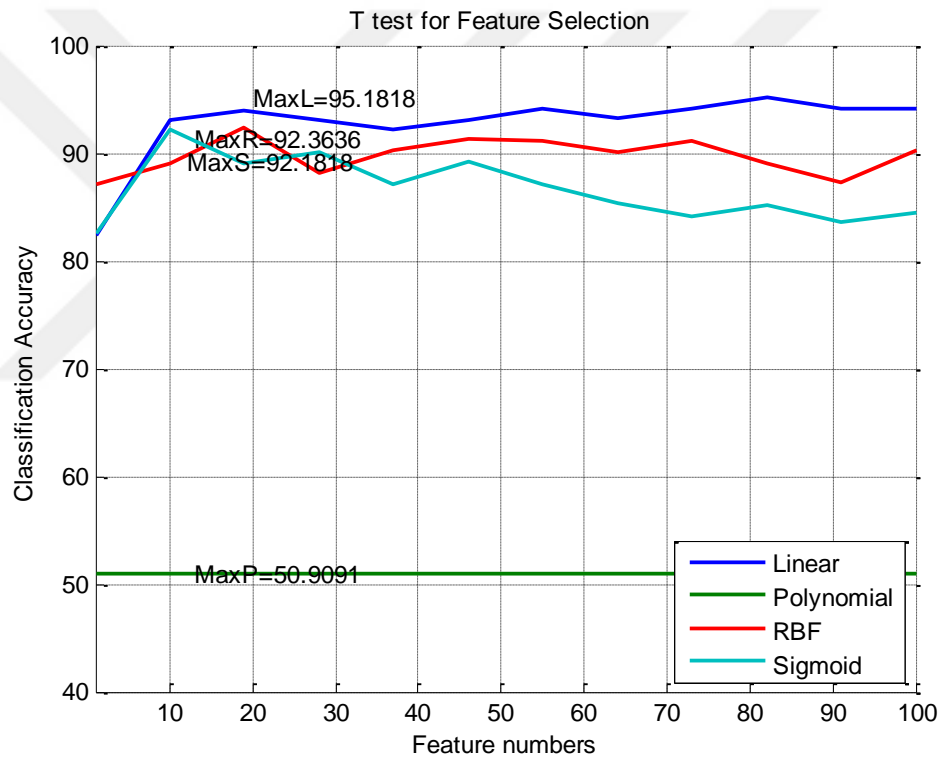


Figure 4.22: Classification performance of SVM for prostate data (normalized)

Table 4.36: CFS for prostate dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-64	Linear	95.18%
1-64	Polynomial	94%
1	RBF	67.72%
1	Sigmoid	50.90%

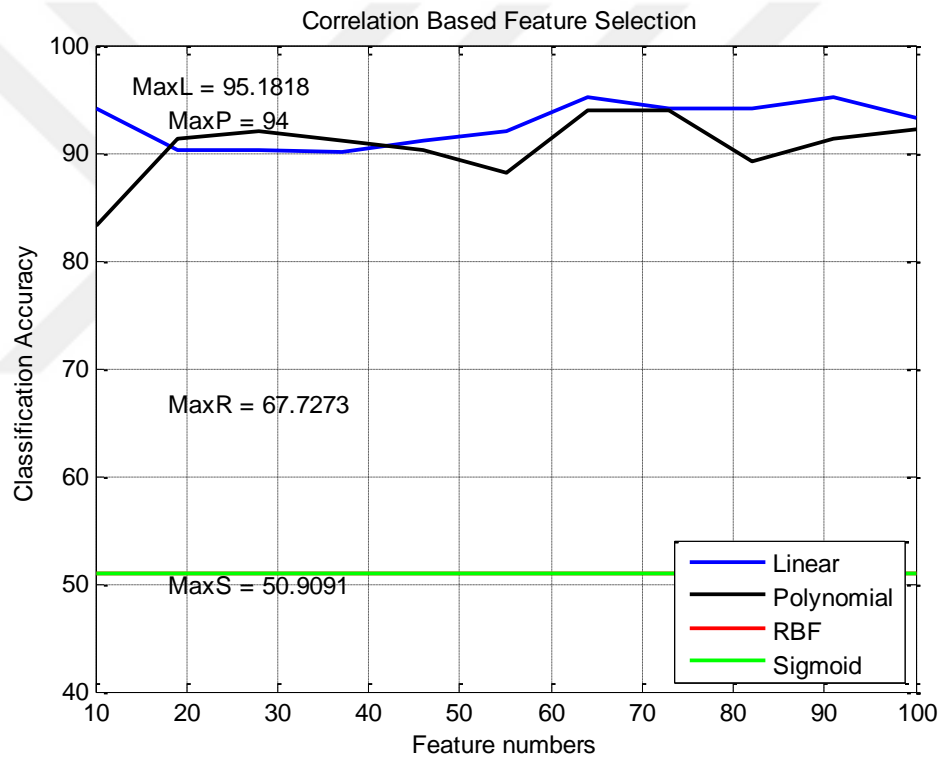


Figure 4.23: Classification performance of SVM for prostate data (raw)

Table 4.37: CFS for prostate dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-19	Linear	95.18%
1	Polynomial	50.90%
1-37	RBF	93.18%
1-19	Sigmoid	85.27%

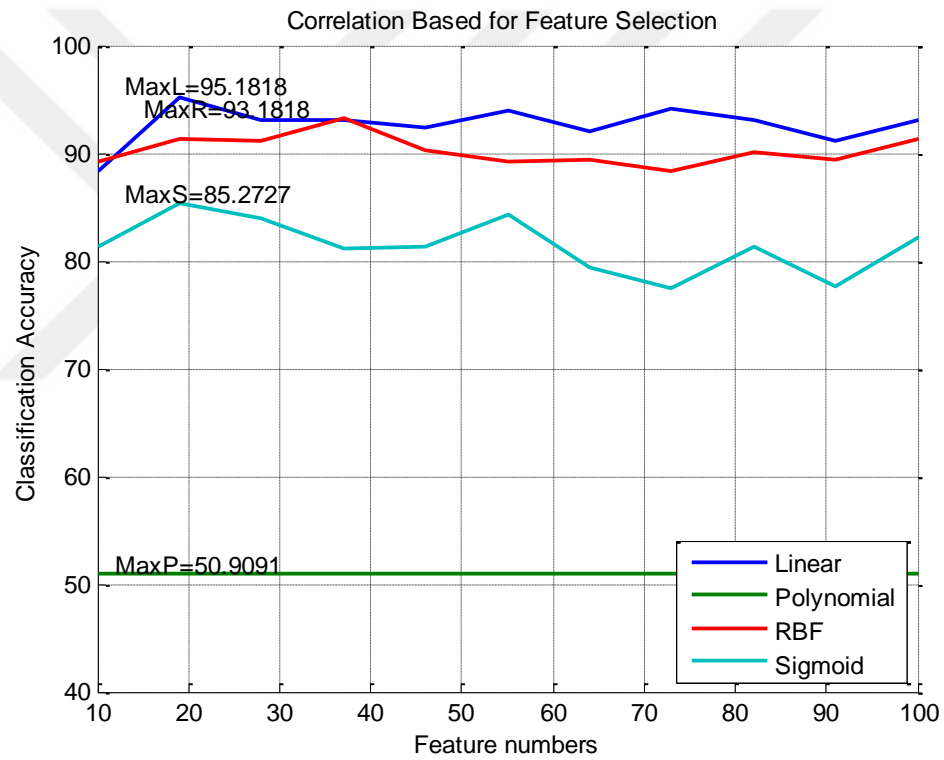


Figure 4.24: Classification performance of SVM for prostate data (normalized)

Table 4.38: Bhattacharyya Distance for prostate dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-82	Linear	91.36%
1-91	Polynomial	87.27%
1	RBF	56.72%
1	Sigmoid	50.90%

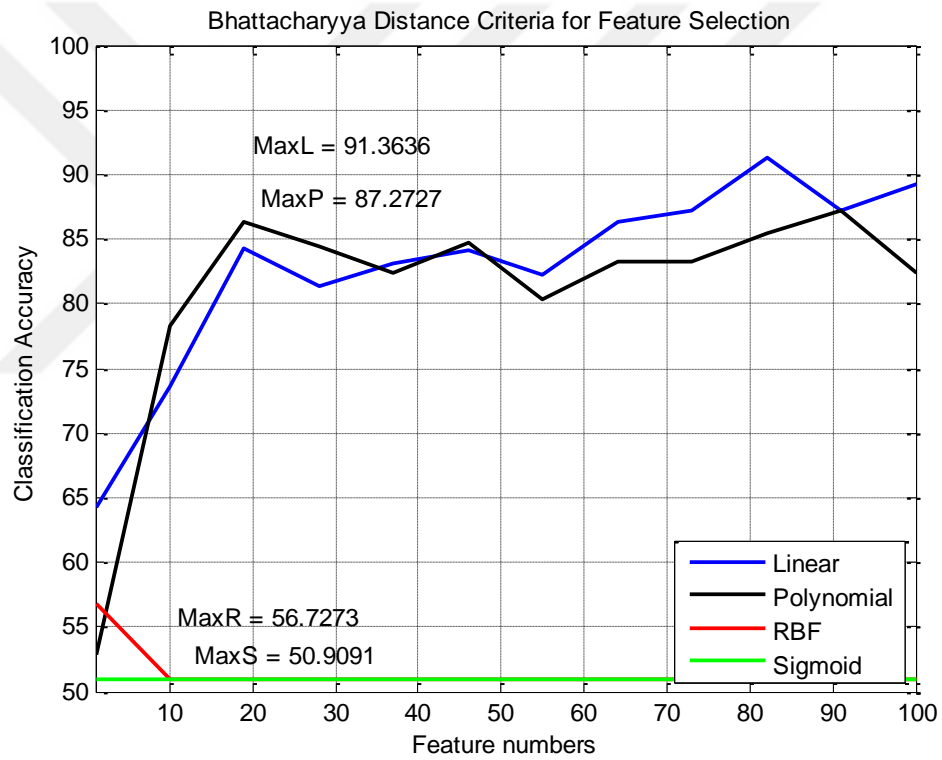


Figure 4.25: Classification performance of SVM for prostate data (raw)

Table 4.39: Bhattacharyya Distance for prostate dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-73	Linear	93.36%
1	Polynomial	50.90%
1-10	RBF	55.90%
1	Sigmoid	50.90%

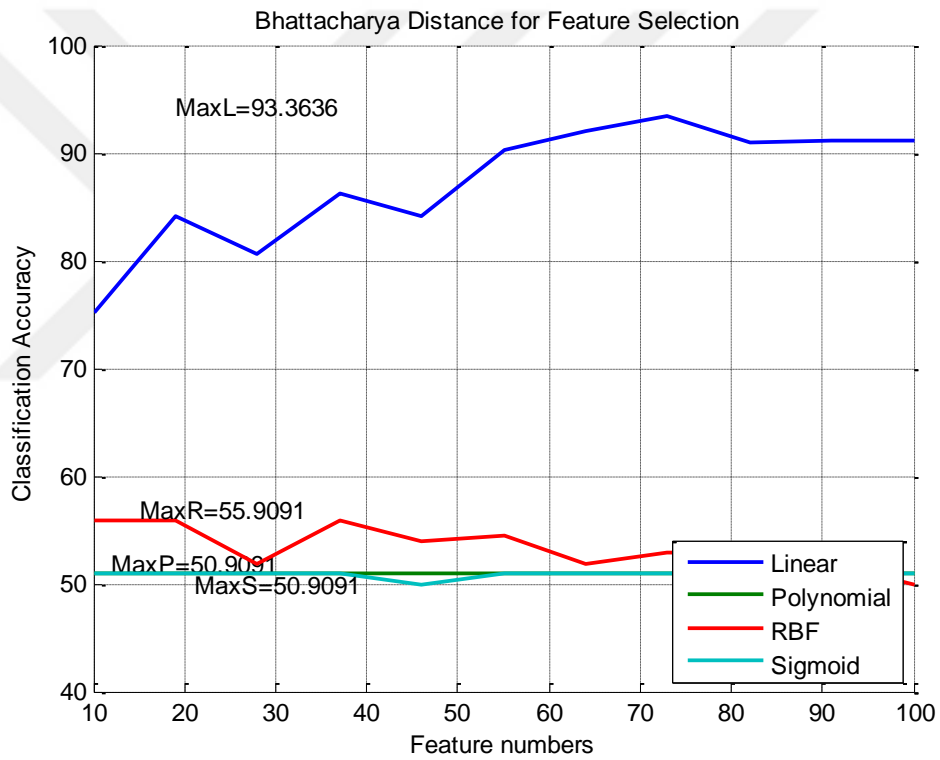


Figure 4.26: Classification performance of SVM for prostate data (normalized)

Table 4.40: Entropy for prostate dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	91.09%
1-73	Polynomial	89.18%
1-10	RBF	50.90%
1	Sigmoid	64.72%

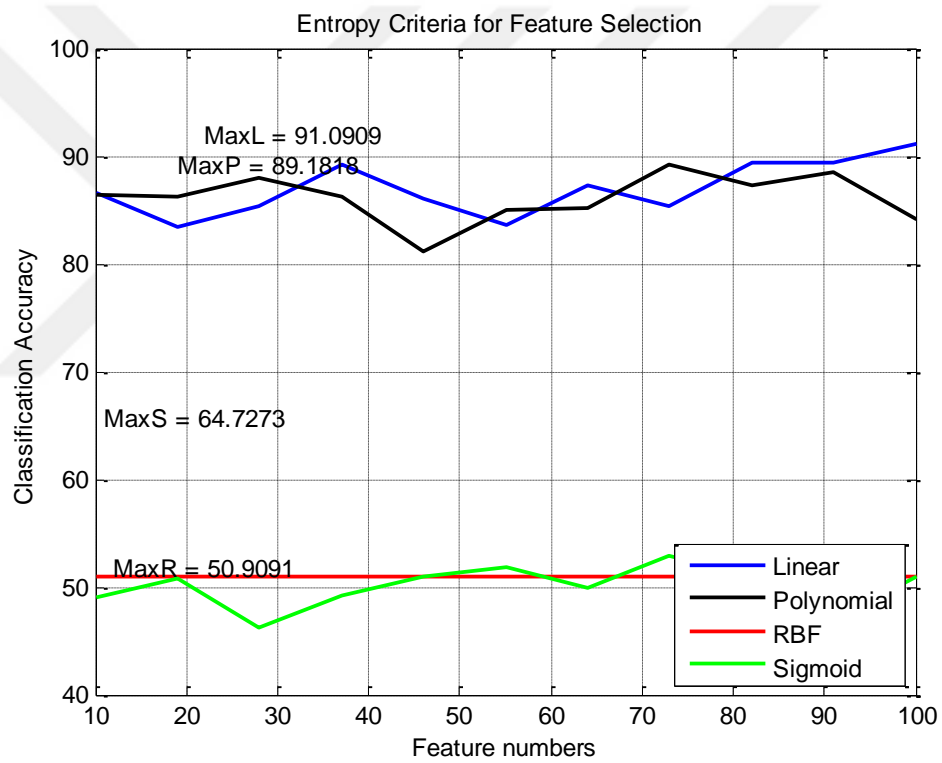


Figure 4.27: Classification performance of SVM for prostate data (raw)

Table 4.41: Entropy for prostate dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	96.18%
1	Polynomial	50.90%
1-37	RBF	69.63%
1-19	Sigmoid	51.90%

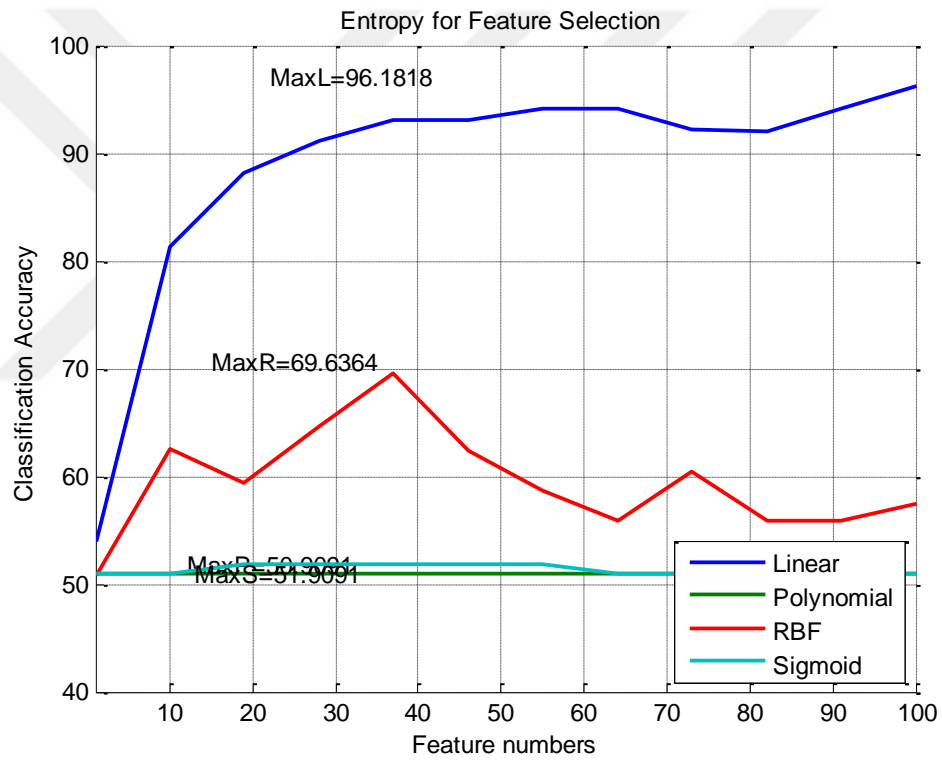


Figure 4.28: Classification performance of SVM for prostate data (normalized)

Table 4.42: ReliefF for prostate dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-91	Linear	92.18%
1-64	Polynomial	92.36%
1	RBF	61%
1-19	Sigmoid	50.90%

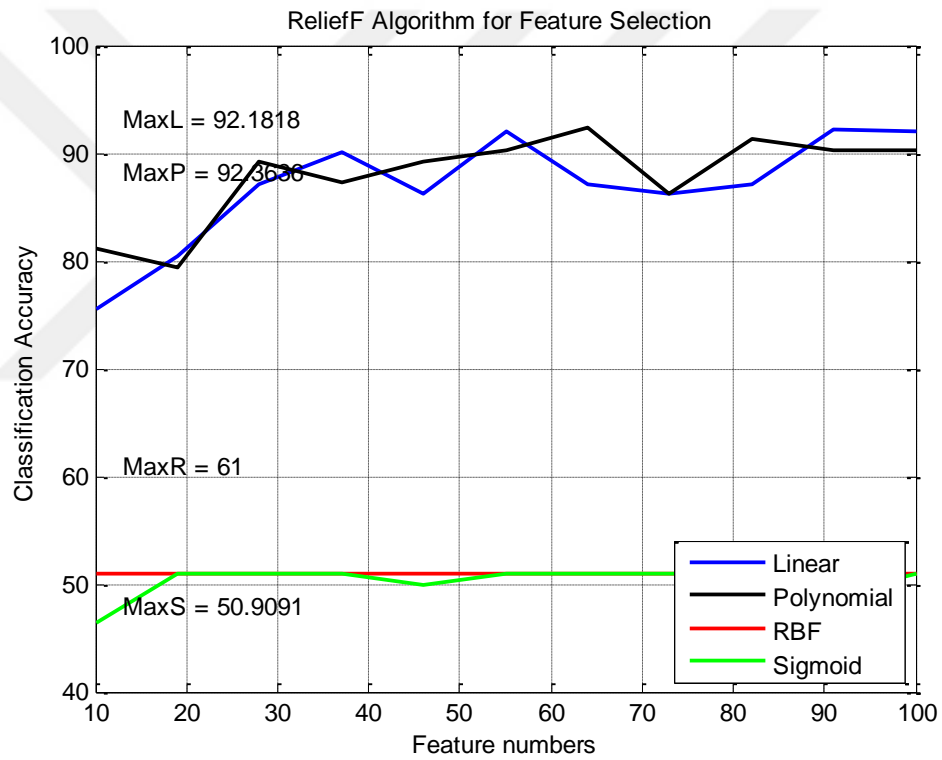


Figure 4.29: Classification performance of SVM for prostate data (raw)

Table 4.43: ReliefF for prostate dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-82	Linear	96.09%
1	Polynomial	50.90%
1-10	RBF	80.36%
1	Sigmoid	65.90%

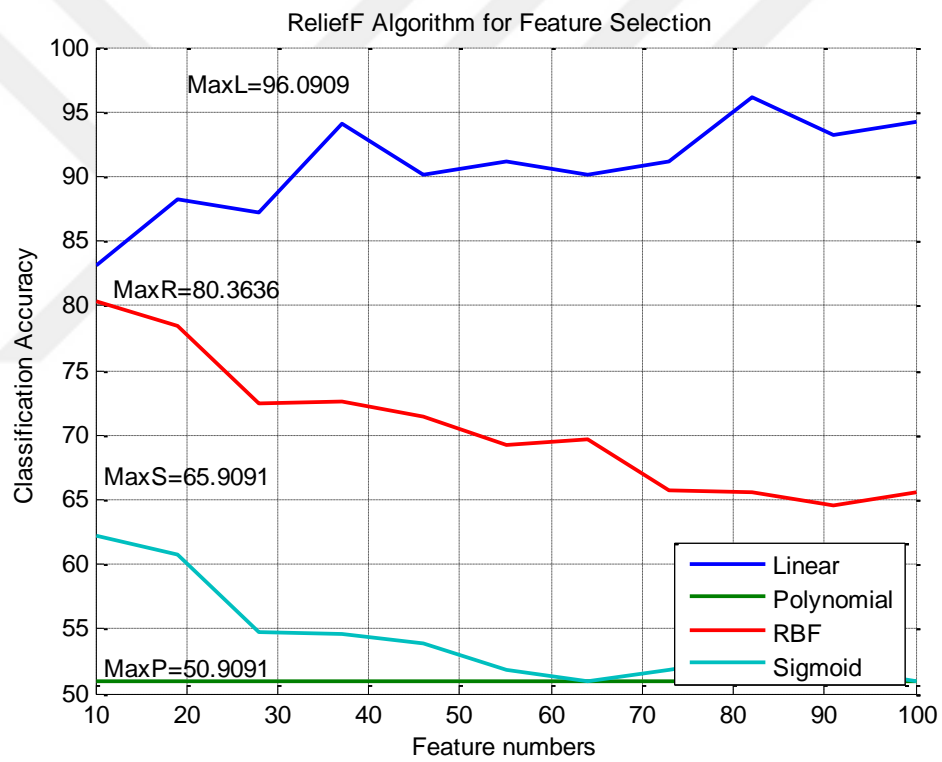


Figure 4.30: Classification performance of SVM for prostate data (normalized)

Table 4.44: Wilcoxon signed-rank test for prostate dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	91.27%
1-100	Polynomial	91%
1	RBF	78.36%
1-10	Sigmoid	52.90%

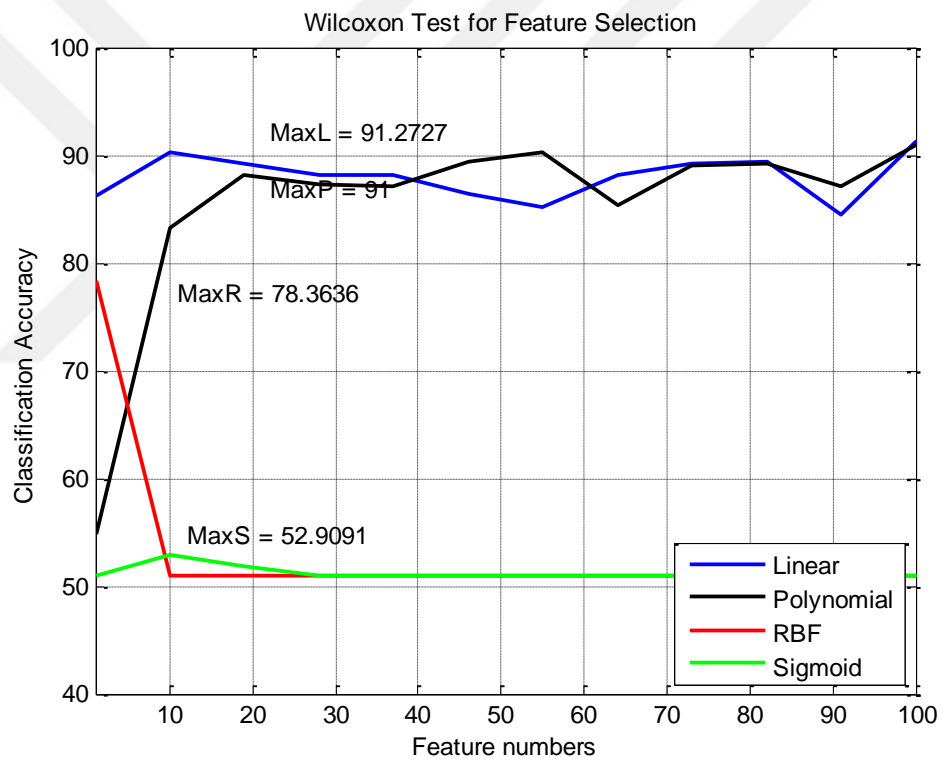


Figure 4.31: Classification performance of SVM for prostate data (raw)

Table 4.45: Wilcoxon signed-rank test for prostate dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-91	Linear	95.09%
1	Polynomial	50.90%
1-10	RBF	93.09%
1-10	Sigmoid	93.18%

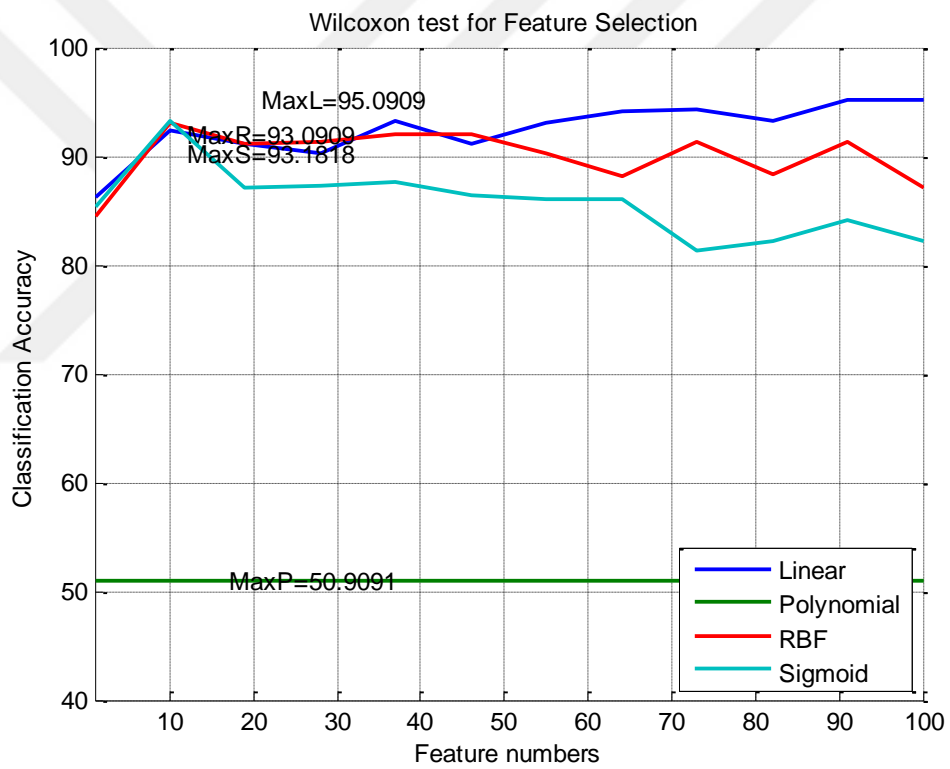


Figure 4.32: Classification performance of SVM for prostate data (normalized)

Table 4.46: mRMR for prostate dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-91	Linear	82.63%
1-100	Polynomial	84.18%
1	RBF	66.45%
1-73	Sigmoid	55.63%

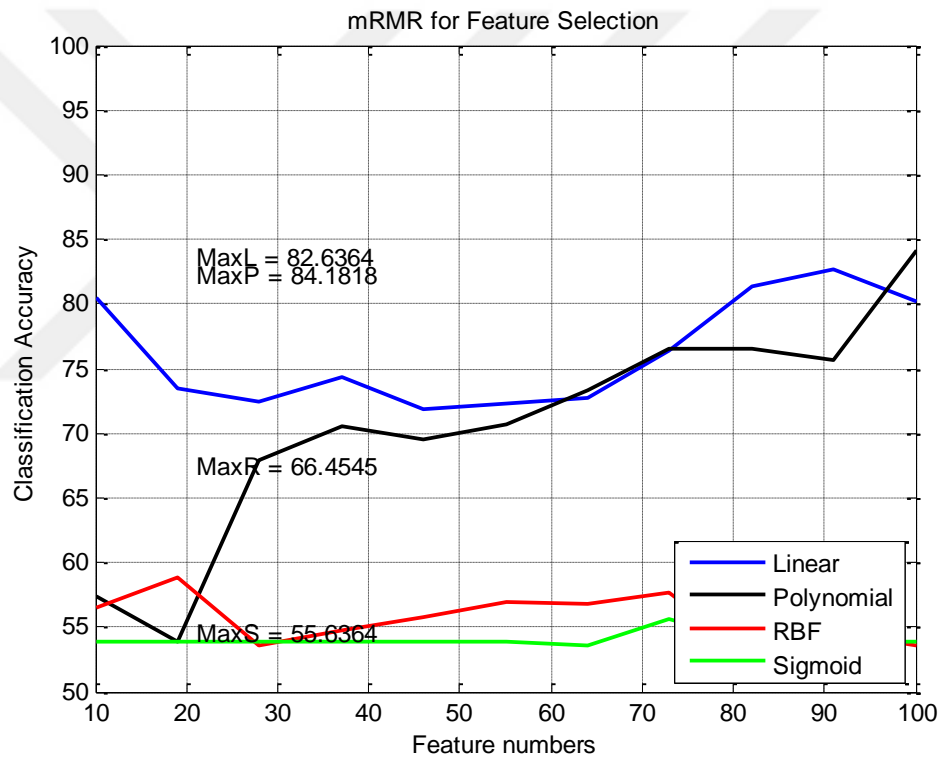


Figure 4.33: Classification performance of SVM for prostate data (raw)

Table 4.47: mRMR for prostate dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	81.36%
1	Polynomial	50.90%
1	RBF	77.36%
1	Sigmoid	73.45%

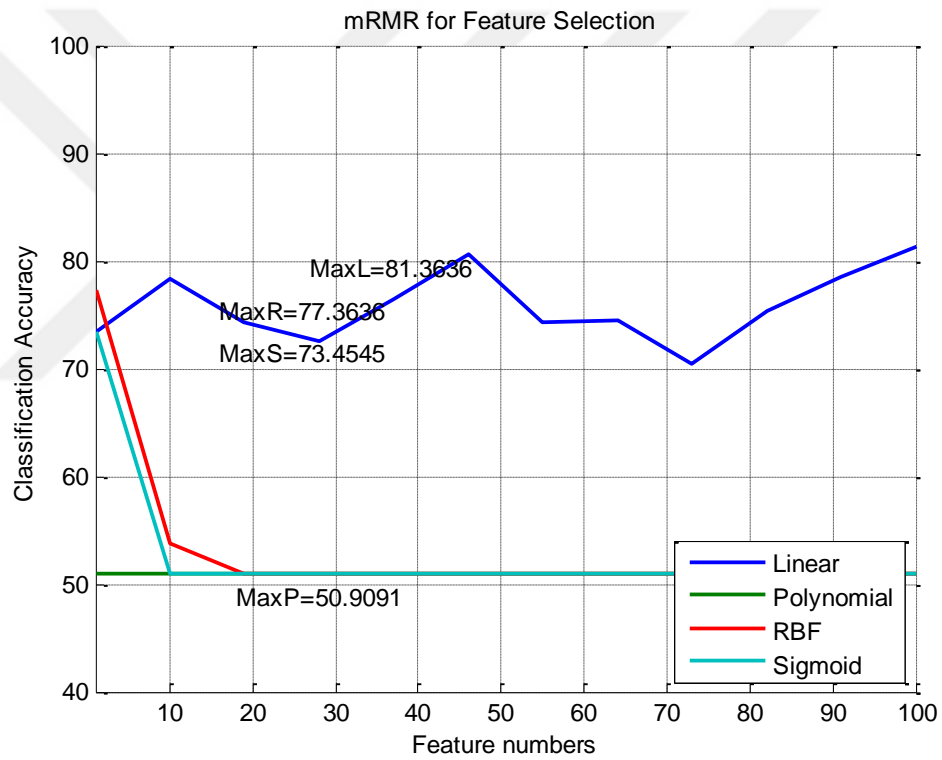


Figure 4.34: Classification performance of SVM for prostate data (normalized)

Table 4.48: DISR for prostate dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	94.27%
1-55	Polynomial	94%
1	RBF	70.36%
1	Sigmoid	53.90%

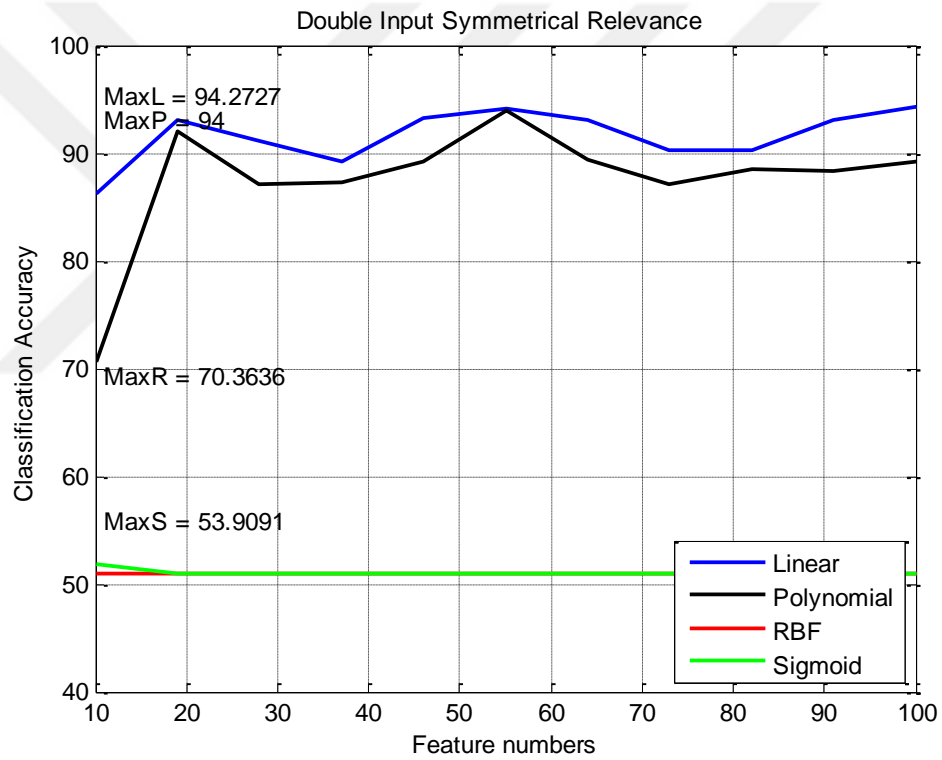


Figure 4.35: Classification performance of SVM for prostate data (raw)

Table 4.49: DISR for prostate dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	95.09%
1	Polynomial	50.90%
1-19	RBF	89.09%
1	Sigmoid	77.27%

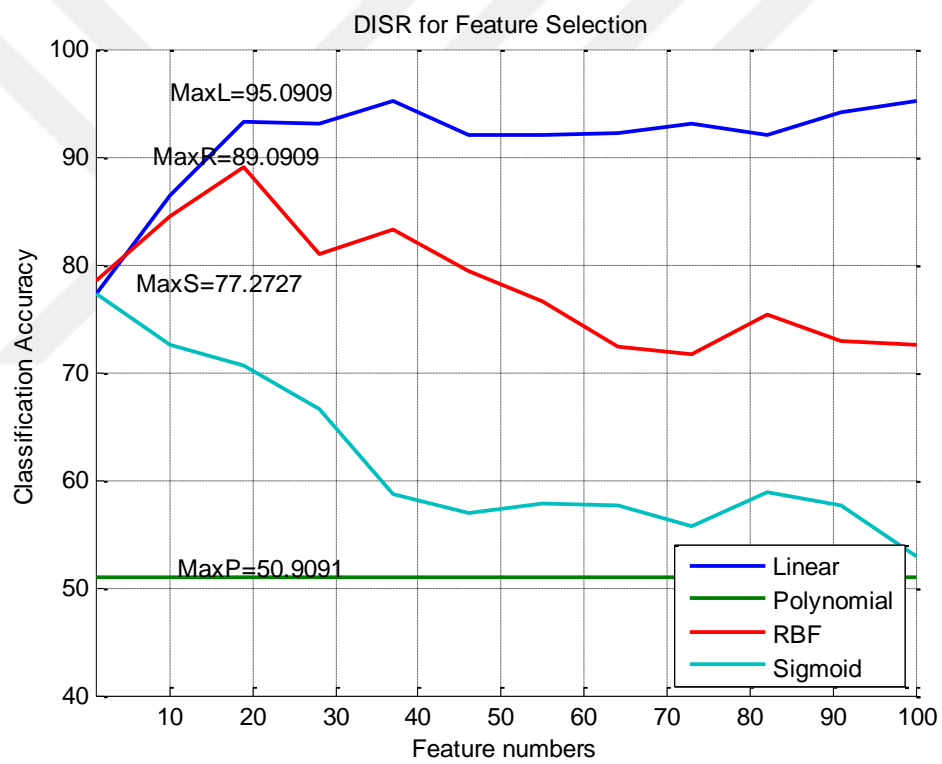


Figure 4.36: Classification performance of SVM for prostate data (normalized)

Table 4.50: ROC for prostate dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-64	Linear	94.18%
1-55	Polynomial	93.09%
1	RBF	81.27%
1-10	Sigmoid	52.90%

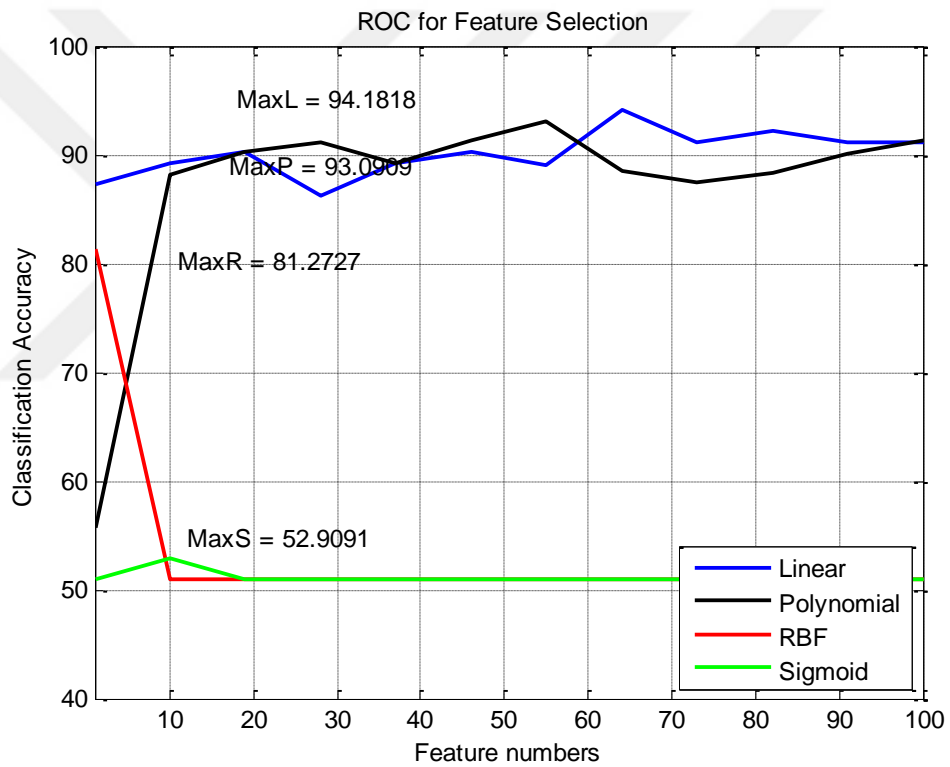


Figure 4.37: Classification performance of SVM for prostate data (raw)

Table 4.51: ROC for prostate dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-64	Linear	95.09%
1	Polynomial	50.90%
1-10	RBF	93.18%
1-19	Sigmoid	92.36%

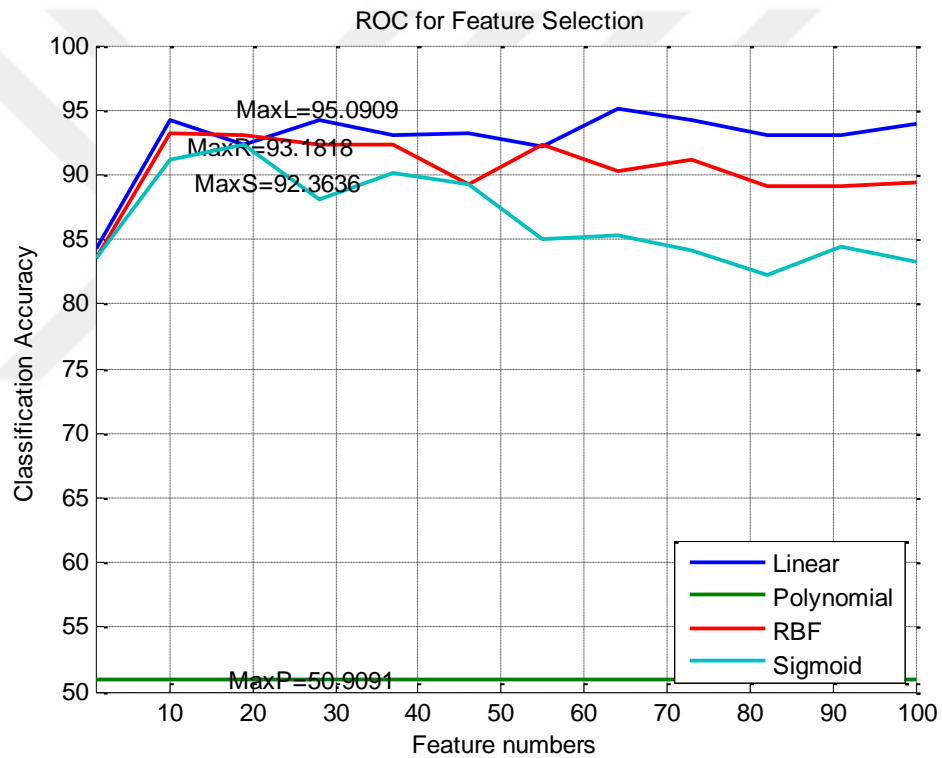


Figure 4.38: Classification performance of SVM for prostate data (normalized)

Table 4.52: Feature numbers for prostate dataset using MLP (raw)

Number of features	Feature Selection Algorithm	Classification Accuracy
1-60	T test	91.18%
1-20	ROC	88.54%
1-40	ReliefF	83.18%
1-20	Wilcoxon	88.45%
1-60	CFS	99.09%
1-30	Entropy	89.36%
1-90	MRMR	93.18%
1-80	DISR	100%
1-10	Bhattacharyya	84.54%

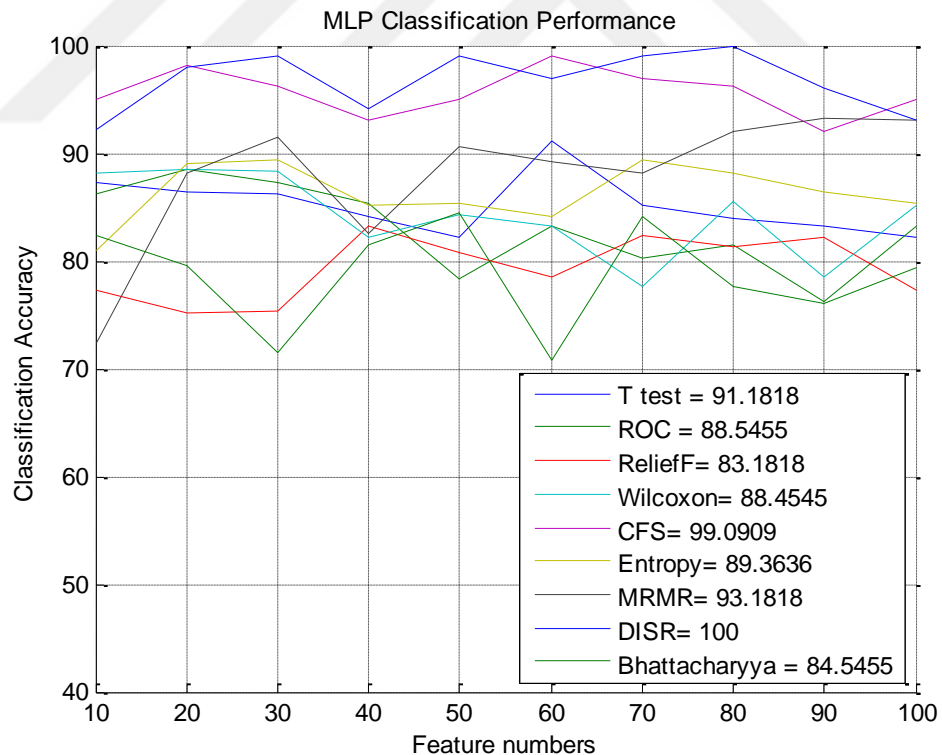


Figure 4.39: Classification performance of MLP for prostate data (raw)

Table 4.53: Feature numbers for prostate dataset using MLP (normalized)

Number of features	Feature Selection Algorithm	Classification Accuracy
1-50	T test	91.27%
1-60	ROC	88.54%
1-90	ReliefF	85.45%
1-50	Wilcoxon	88.45%
1-10	CFS	94.09%
1-50	Entropy	84.72%
1-90	MRMR	87.27%
1-30	DISR	97%
1-30	Bhattacharyya	79.18%

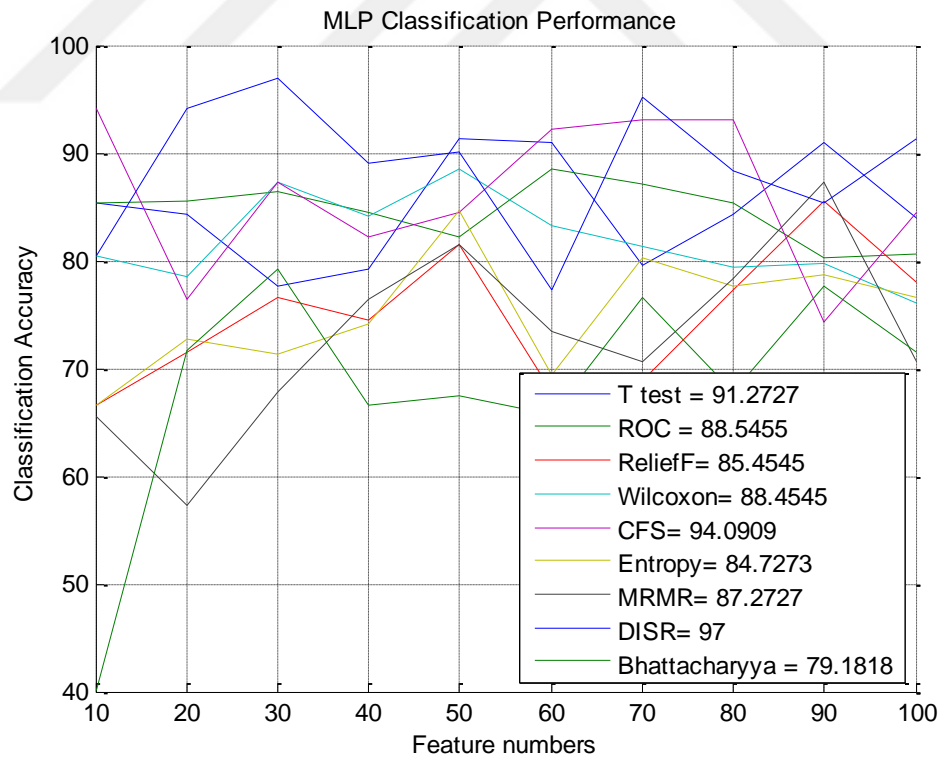


Figure 4.40: Classification performance of MLP for prostate data (normalized)

4.3 Experimental Results for DLBCL Data

Same experimental procedure followed for the DLBCL data. Nine different feature selection methods applied to find an optimal gene subset. Sufficiency of the gene subsets for classification evaluated using SVM and multi-layer perceptron. 10-fold cross validation is used to separate the data into train and test sets. C SVM with four different kernels is applied to determine the classification performance. Feed forward back-propagation perceptron with four layers which has Levenberg-Marquadt as training and mean squared error as performance function is used to evaluate the classification performance. Feature number for the gene subsets is selected according to previous studies and adjusted in the range of [1 100]. Different gene subsets constituted with 9 feature increments and the classifiers evaluated over these 10 different gene subsets. First, the classification performed with the raw data. Normalization so that the data will have 0 mean, 1 standard deviation and scaling in the range of [-1 1] applied to see its effect on gene selection and classification performance.

Table 4.54: T test statistic feature selection for DLBCL dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-91	Linear	93.33%
1-82	Polynomial	88.57%
1	RBF	75.47%
1	Sigmoid	75.47%

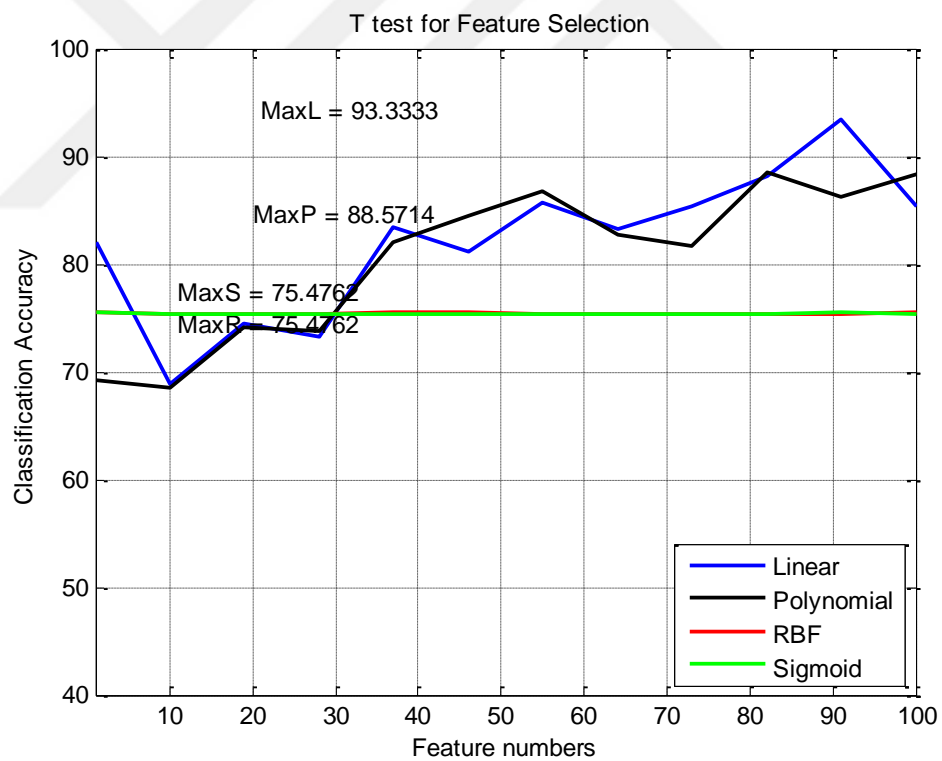


Figure 4.41: Classification performance of SVM for DLBCL data (raw)

Table 4.55: T test statistic feature selection for DLBCL dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-73	Linear	88.39%
1	Polynomial	75.47%
1-19	RBF	76.78%
1-91	Sigmoid	75.47%

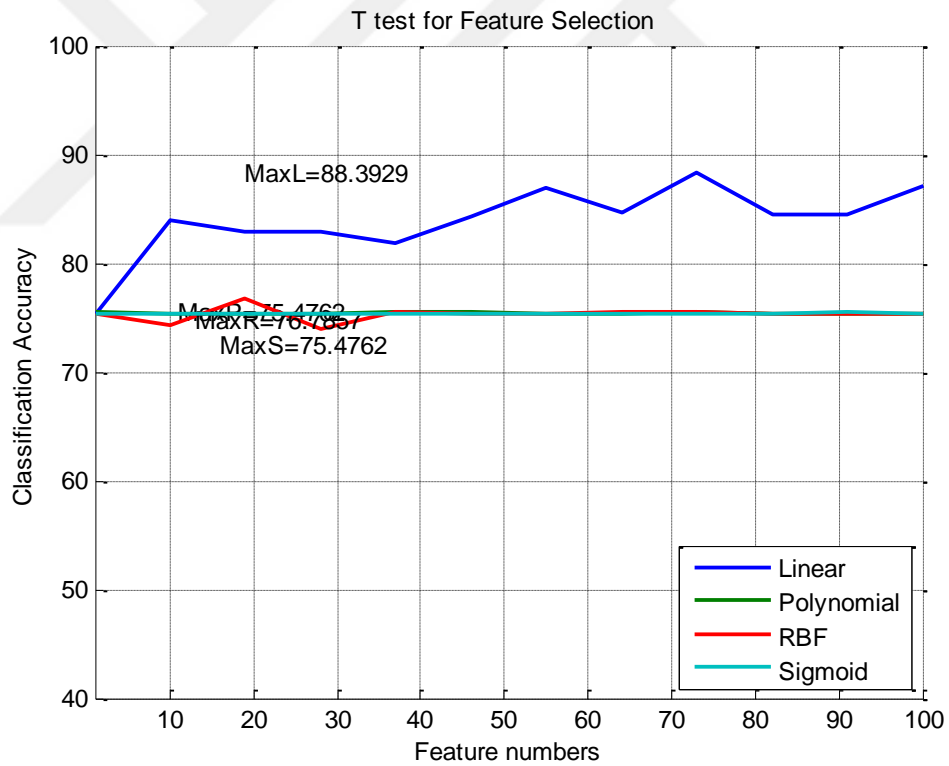


Figure 4.42: Classification performance of SVM for DLBCL data (normalized)

Table 4.56: CFS for DLBCL dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-64	Linear	95%
1-55	Polynomial	100%
1	RBF	76.42%
1	Sigmoid	75.47%

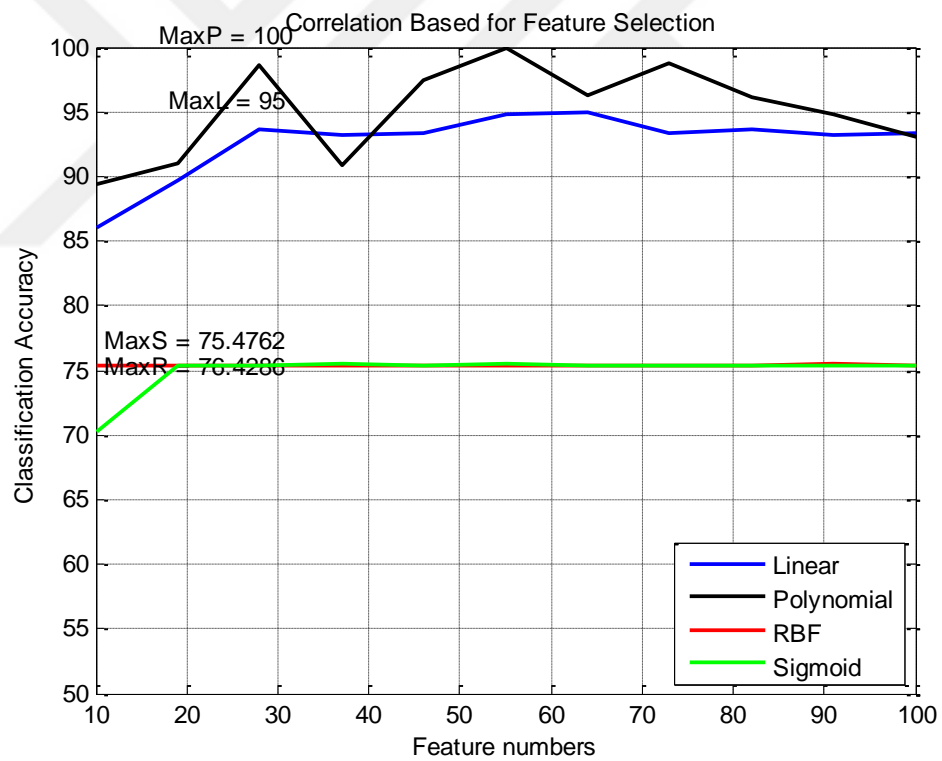


Figure 4.43: Classification performance of SVM for DLBCL data (raw)

Table 4.57: CFS for DLBCL dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-73	Linear	93.39%
1	Polynomial	75.47%
1-10	RBF	80.89%
1	Sigmoid	75.47%

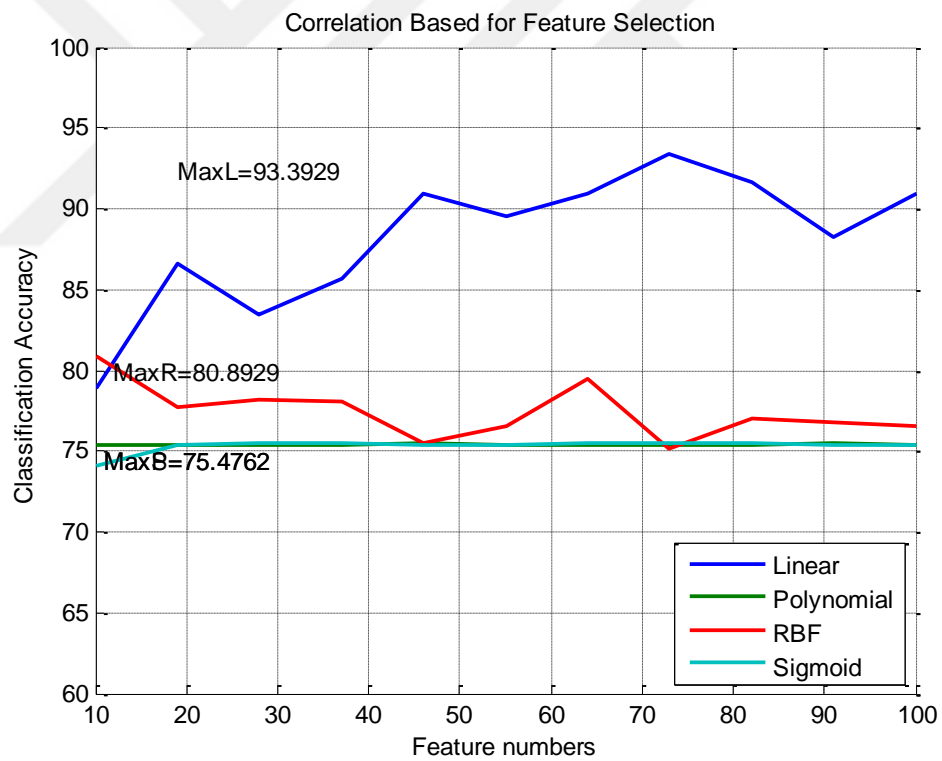


Figure 4.44: Classification performance of SVM for DLBCL data (normalized)

Table 4.58: Bhattacharyya Distance for DLBCL dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-55	Linear	93.75%
1-64	Polynomial	95%
1-46	RBF	75.47%
1-46	Sigmoid	75.47%

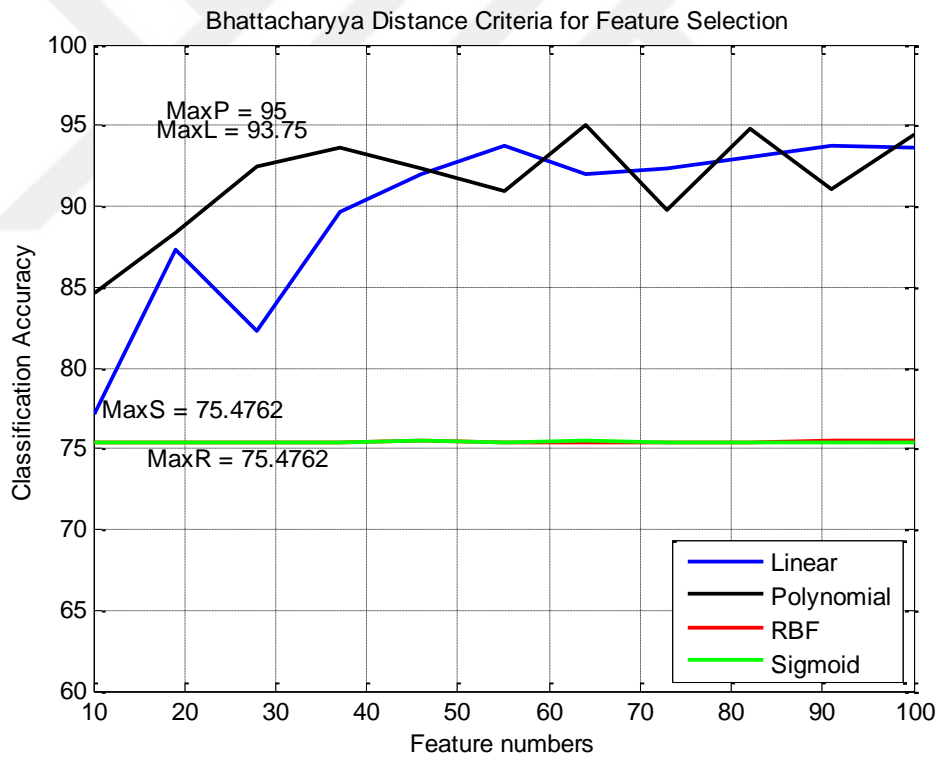


Figure 4.45: Classification performance of SVM for DLBCL data (raw)

Table 4.59: Bhattacharyya Distance for DLBCL dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-37	Linear	83.57%
1-19	Polynomial	75.47%
1-19	RBF	75.47%
1-64	Sigmoid	75.47%

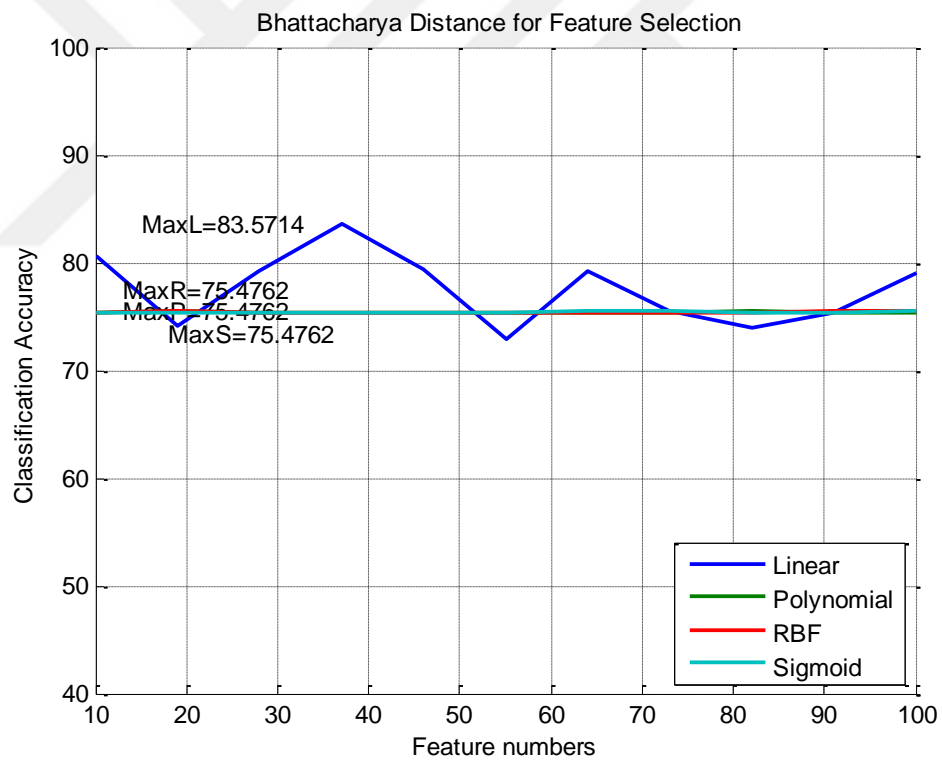


Figure 4.46: Classification performance of SVM for DLBCL data (normalized)

Table 4.60: Entropy for DLBCL dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-82	Linear	92.32%
1-37	Polynomial	87.14%
1-19	RBF	75.47%
1-28	Sigmoid	79.46%

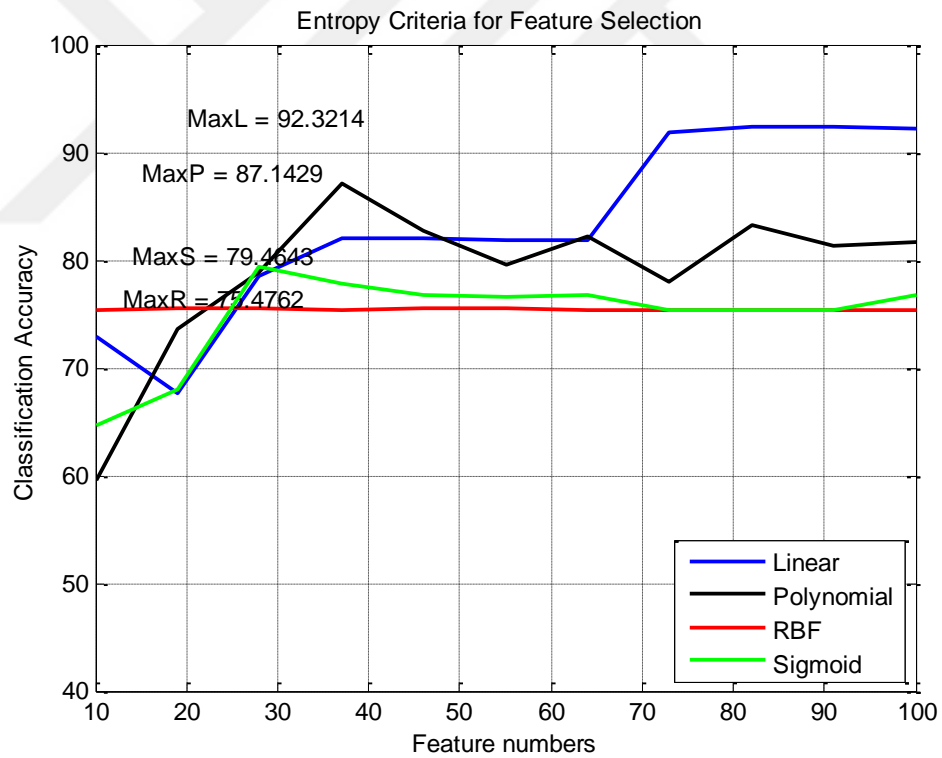


Figure 4.47: Classification performance of SVM for DLBCL data (raw)

Table 4.61: Entropy for DLBCL dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	81.60%
1-55	Polynomial	75.47%
1-28	RBF	75.47%
1-28	Sigmoid	75.47%

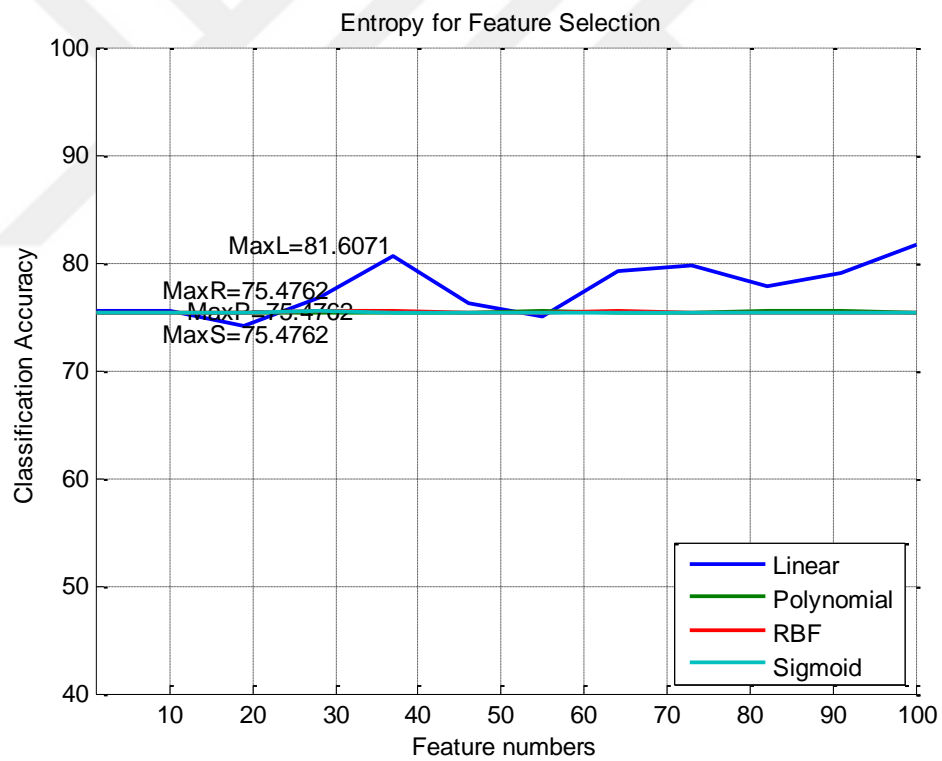


Figure 4.48: Classification performance of SVM for DLBCL data (normalized)

Table 4.62: ReliefF for DLBCL dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-37	Linear	96.07%
1-100	Polynomial	94.82%
1	RBF	75.47%
1-91	Sigmoid	75.47%

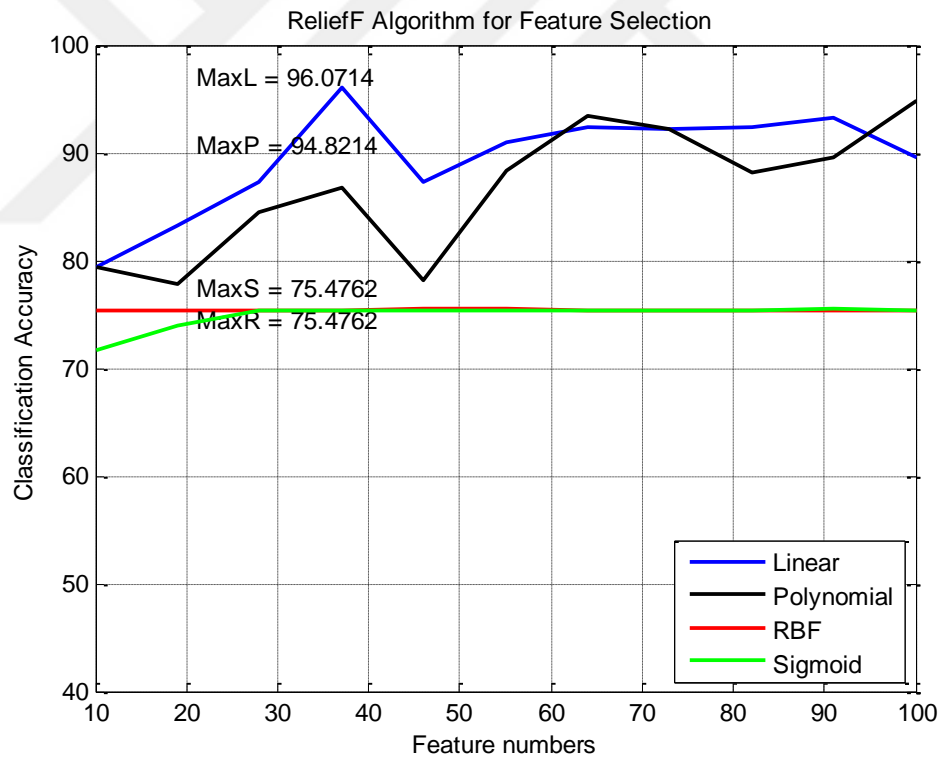


Figure 4.49: Classification performance of SVM for DLBCL data (raw)

Table 4.63: ReliefF for DLBCL dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	92.14%
1	Polynomial	75.47%
1-10	RBF	76.60%
1-37	Sigmoid	75.47%

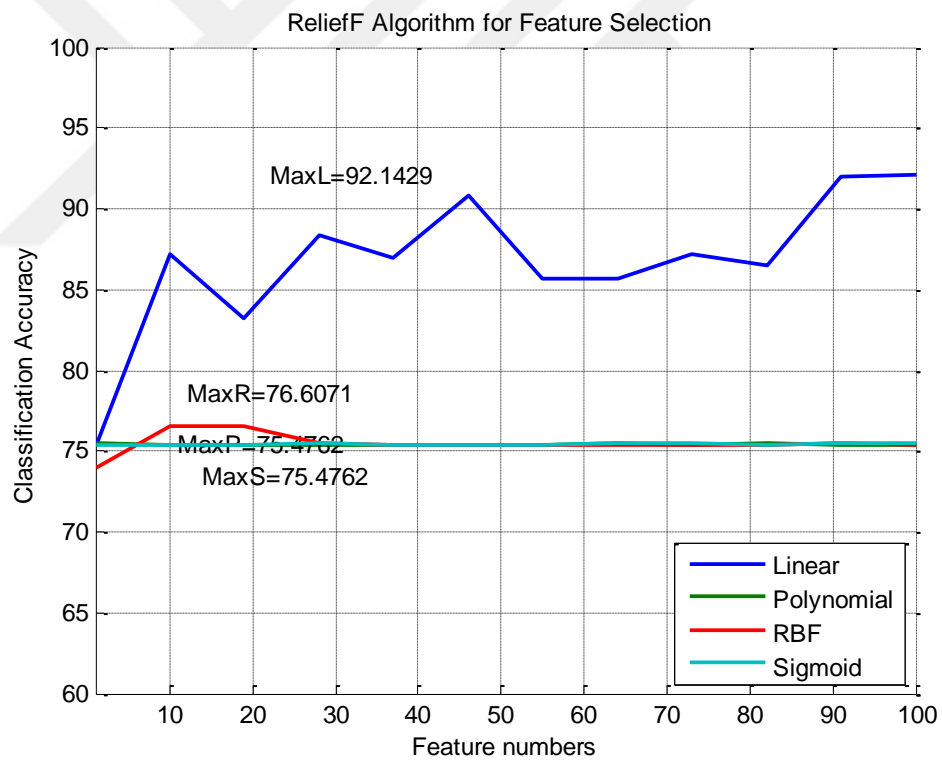


Figure 4.50: Classification performance of SVM for DLBCL data (normalized)

Table 4.64: Wilcoxon signed-rank test for DLBCL dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-37	Linear	92.32%
1-100	Polynomial	93.21%
1-91	RBF	75.47%
1	Sigmoid	75.47%

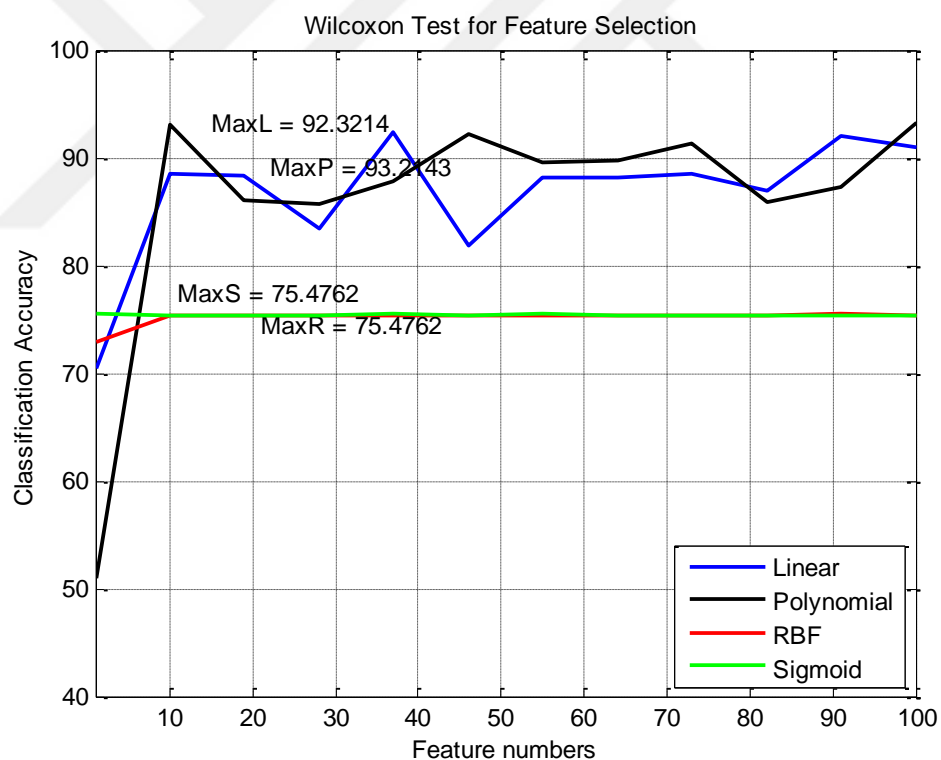


Figure 4.51: Classification performance of SVM for DLBCL data (raw)

Table 4.65: Wilcoxon signed-rank test for DLBCL dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-37	Linear	94.64%
1	Polynomial	75.47%
1-19	RBF	89.64%
1	Sigmoid	75.47%

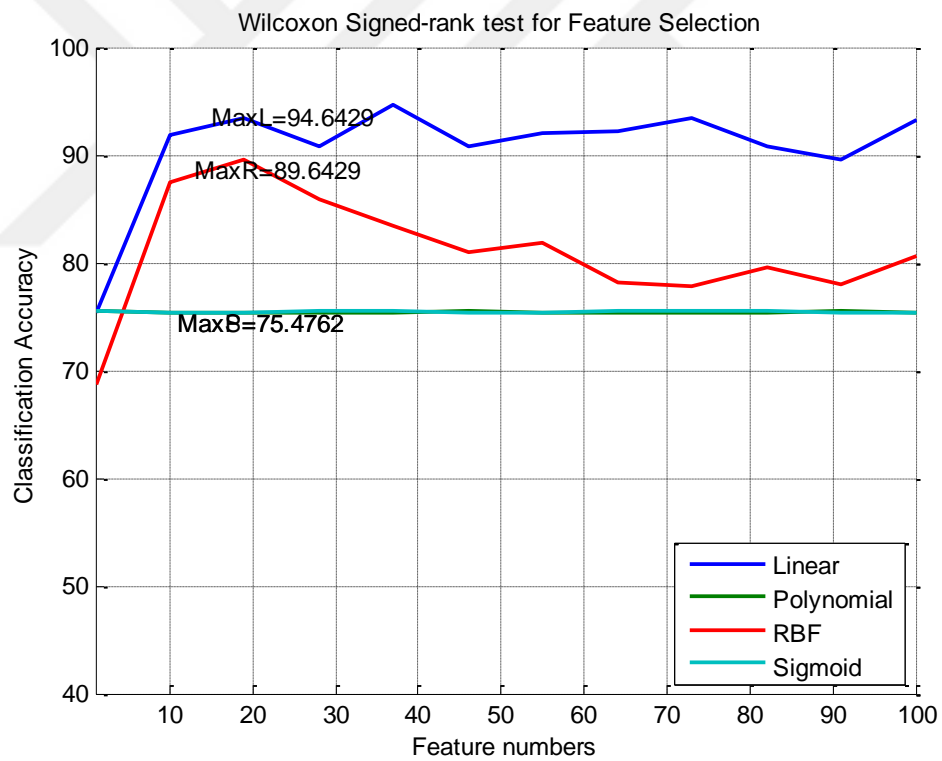


Figure 4.52: Classification performance of SVM for DLBCL data (normalized)

Table 4.66: MRMR for DLBCL dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-46	Linear	96.25%
1-37	Polynomial	93.75%
1	RBF	77.85%
1	Sigmoid	75.47%

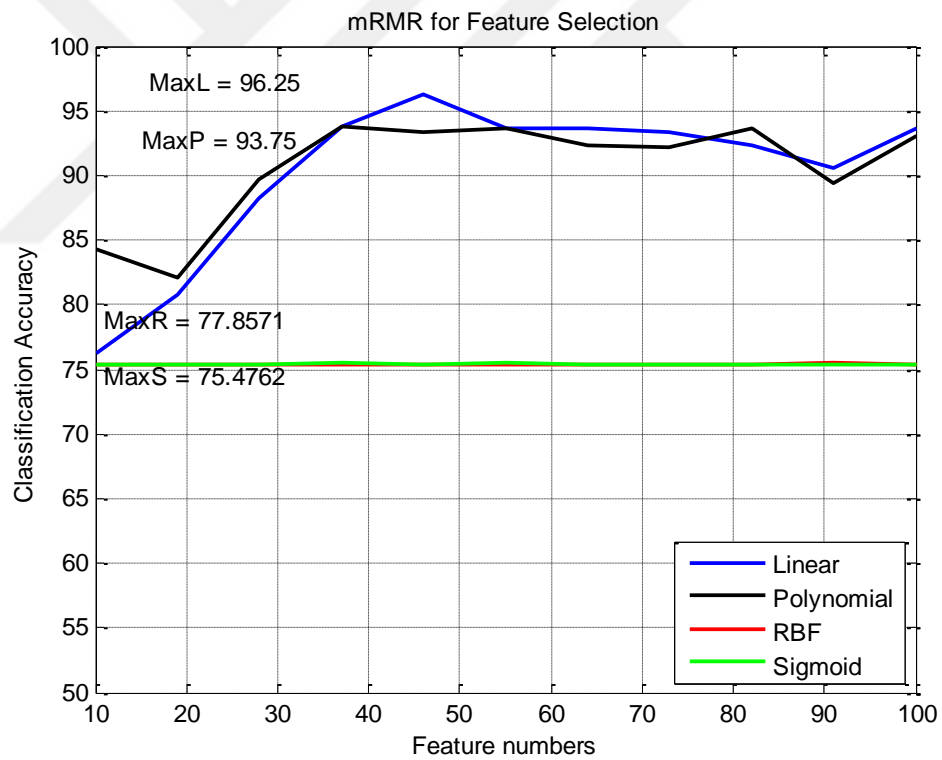


Figure 4.53: Classification performance of SVM for DLBCL data (raw)

Table 4.67: MRMR for DLBCL dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-73	Linear	90.89%
1	Polynomial	75.47%
1	RBF	75.47%
1	Sigmoid	75.47%

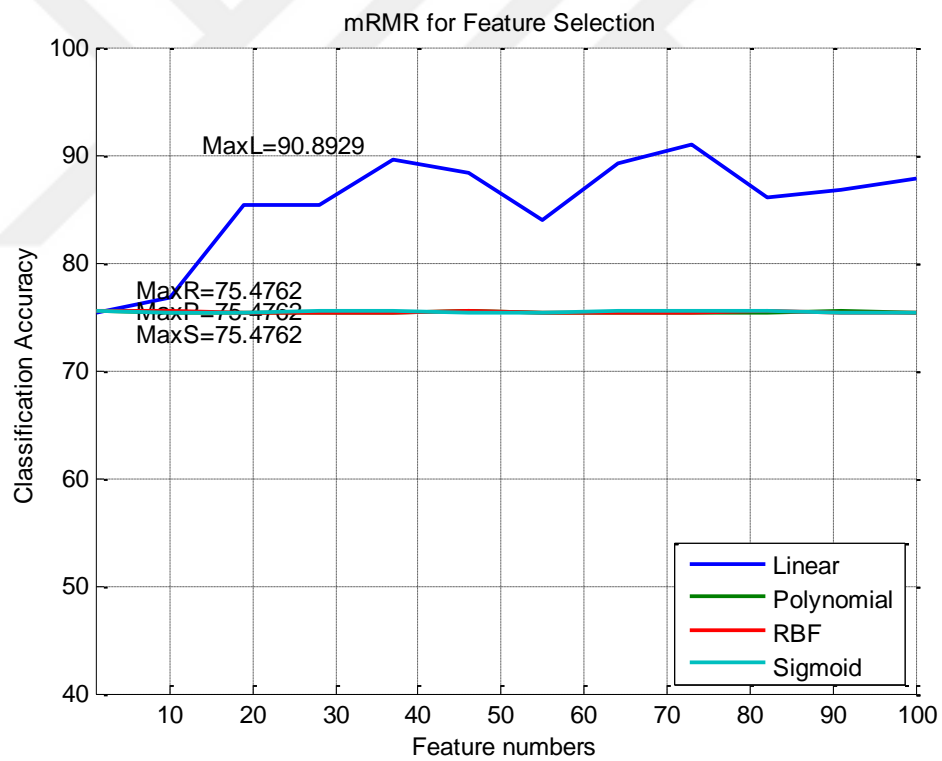


Figure 4.54: Classification performance of SVM for DLBCL data (normalized)

Table 4.68: DISR for DLBCL dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	91.01%
1-82	Polynomial	87.14%
1-91	RBF	75.47%
1	Sigmoid	75.47%

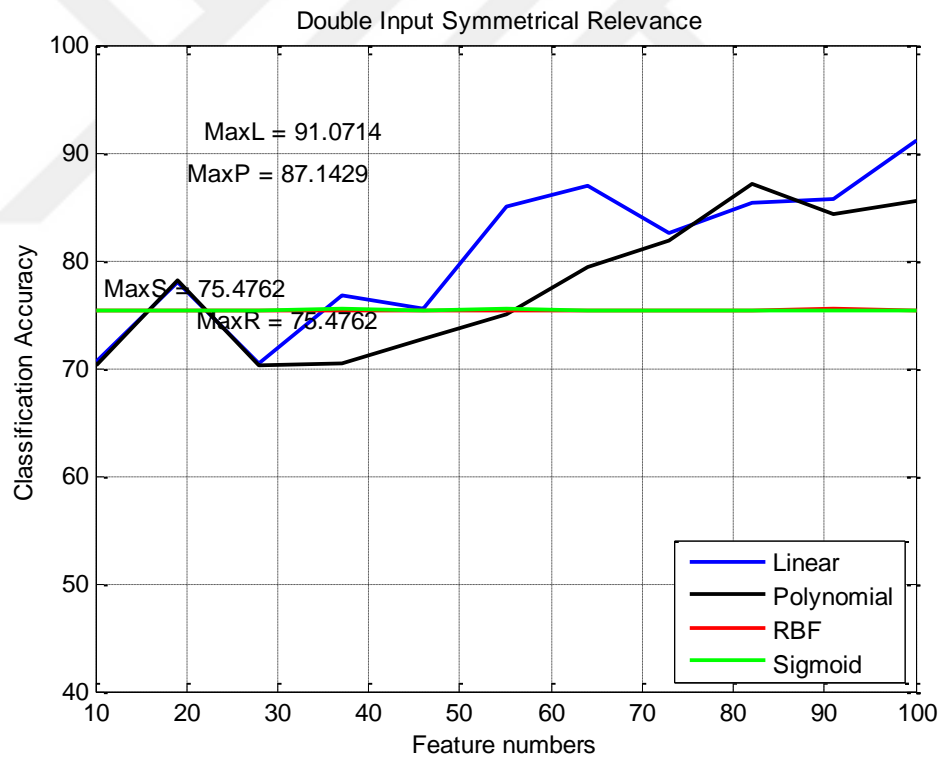


Figure 4.55: Classification performance of SVM for DLBCL data (raw)

Table 4.69: DISR for DLBCL dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-100	Linear	88.21%
	Polynomial	75.47%
1	RBF	75.47%
1	Sigmoid	75.47%

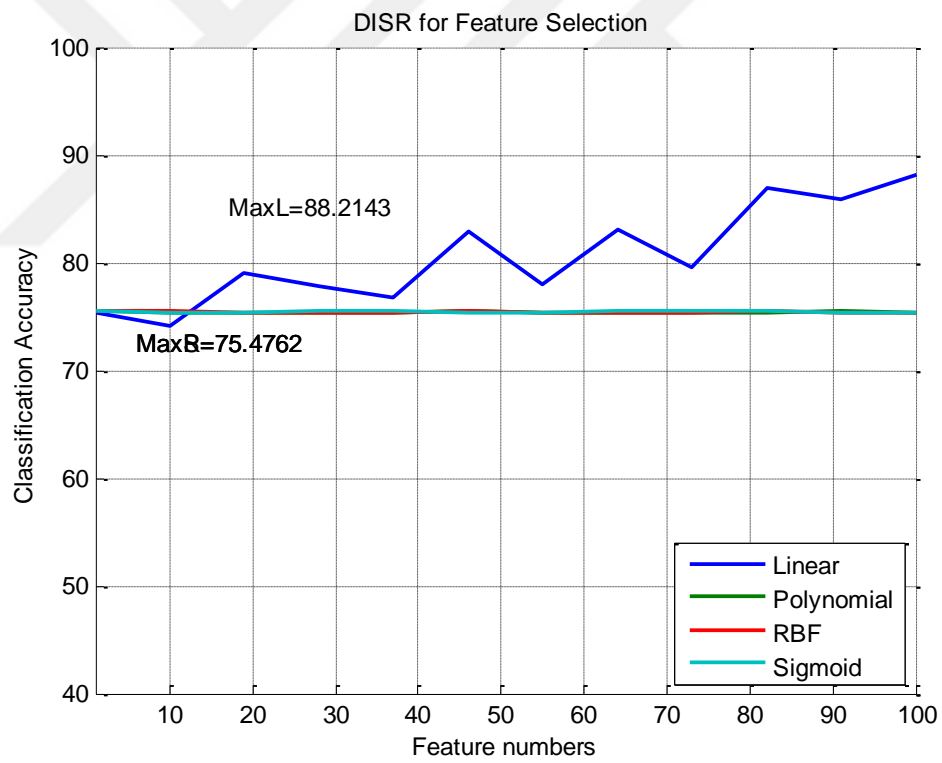


Figure 4.56: Classification performance of SVM for DLBCL data (normalized)

Table 4.70: ROC for DLBCL dataset (raw)

Number of features	Kernel Type	Classification Accuracy
1-55	Linear	90.71%
1-73	Polynomial	84.82%
1-91	RBF	75.47%
1	Sigmoid	75.47%

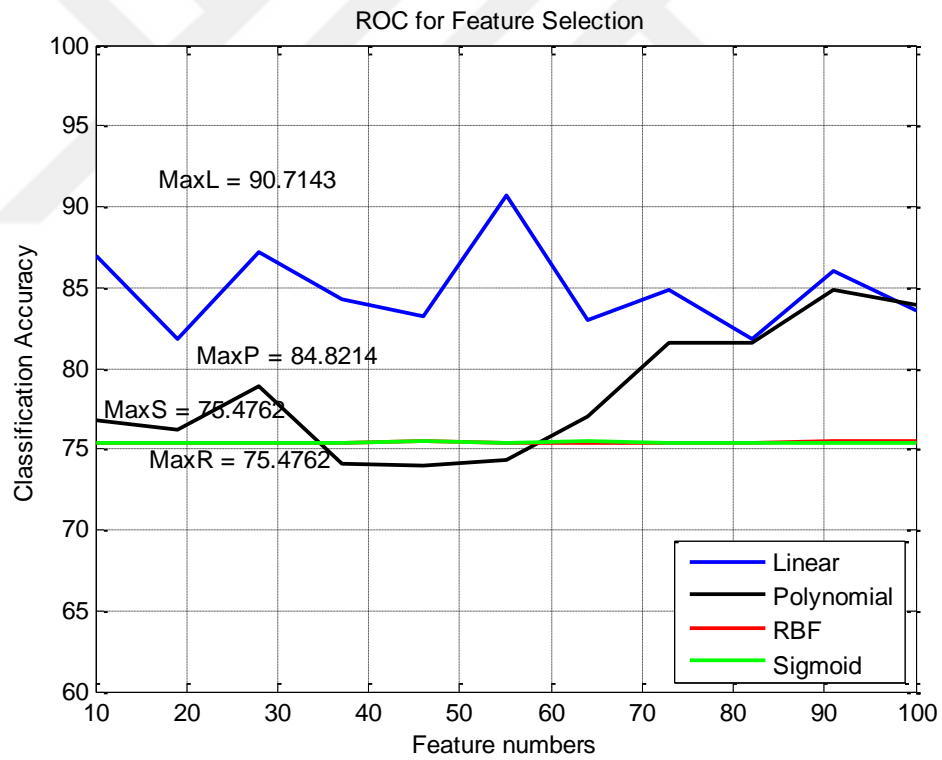


Figure 4.57: Classification performance of SVM for DLBCL data (raw)

Table 4.71: ROC for DLBCL dataset (normalized)

Number of features	Kernel Type	Classification Accuracy
1-82	Linear	89.58%
1-19	Polynomial	75.47%
1-91	RBF	75.47%
1-64	Sigmoid	75.47%

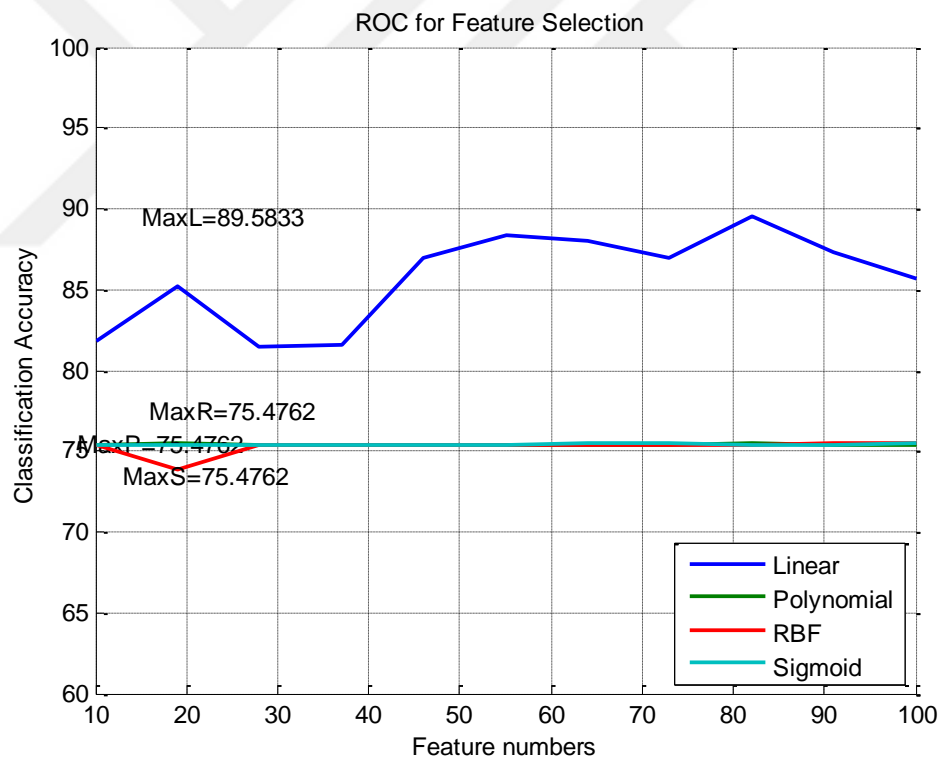


Figure 4.58: Classification performance of SVM for DLBCL data (normalized)

Table 4.72: Feature numbers for DLBCL dataset using MLP (raw)

Number of features	Feature Selection Algorithm	Classification Accuracy
1-70	T test	89.82%
1-90	ROC	86.96%
1-90	ReliefF	83.75%
1-100	Wilcoxon	89.82%
1-90	CFS	100%
1-90	Entropy	89.28%
1-90	MRMR	98.75%
1-100	DISR	97.5%
1-50	Bhattacharyya	95%

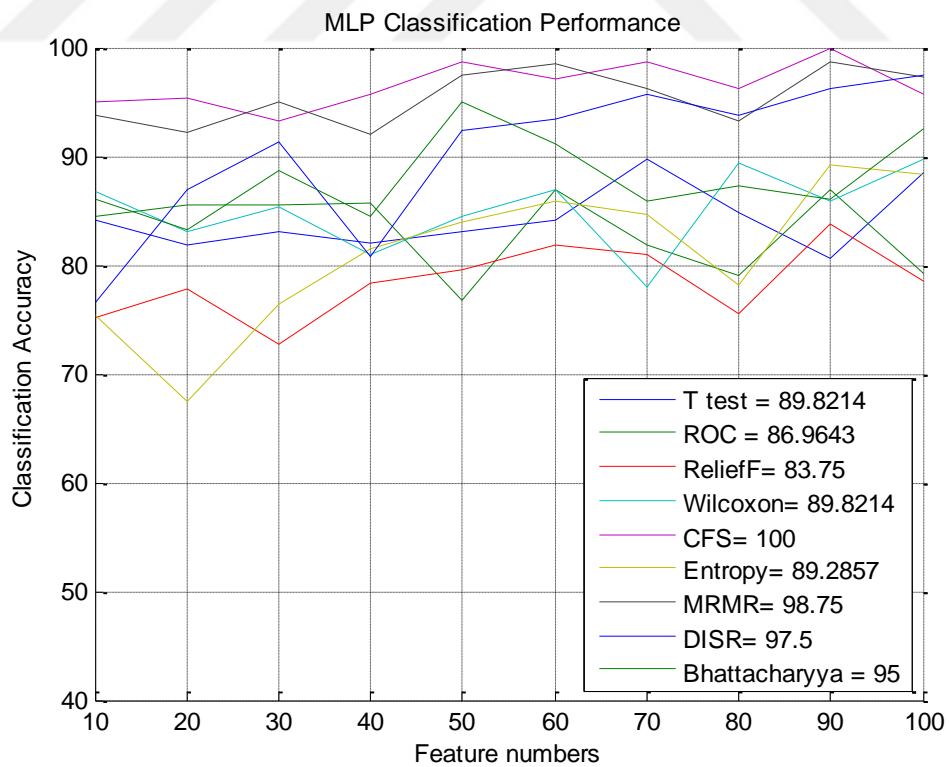


Figure 4.59: Classification performance of MLP for DLBCL data (raw)

Table 4.73: Feature numbers for DLBCL dataset using MLP (normalized)

Number of features	Feature Selection Algorithm	Classification Accuracy
1-90	T test	85.35%
1-10	ROC	83.21%
1-60	ReliefF	83.03%
1-90	Wilcoxon	89.82%
1-60	CFS	97.32%
1-100	Entropy	70.53%
1-60	MRMR	95%
1-80	DISR	93.39%
1-80	Bhattacharyya	66.42%

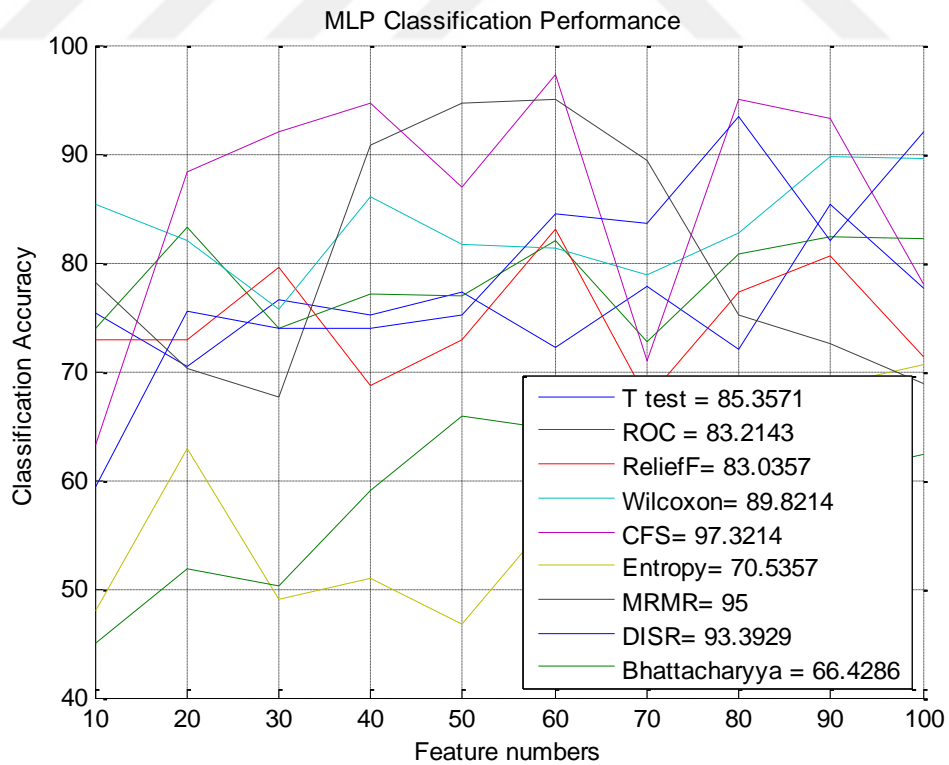


Figure 4.60: Classification performance of MLP for DLBCL data (normalized)

The performance of the above feature selection algorithms and classifiers are evaluated according to their classification performance and required feature/gene numbers. There is a trade-off between these two factors. Although, high classification accuracy is desired and important for the performance of the algorithm, required feature number should be small. Previous studies show that less than 50 features are enough to classify cancer types with good classification performance [40], [44].

For the leukemia data classification problem, all of the algorithms performed well, above 80% classification accuracy. When the raw data used for classification, t test as feature selection algorithm with linear SVM gave the best result; 100% classification accuracy and used only 10 features/genes to classify. Normalization and scaling for leukemia data increased the classification performance but affected the required feature number negatively for most of the feature selection algorithms. ROC as feature selection algorithms with linear SVM gave the best result for normalized and scaled leukemia data; 98.75% classification accuracy with only 10 features required. When multi-layer perceptron is used for classification, CFS gave the best result with 100% accuracy, when raw data is used and required only 10 features. Normalization and scaling of the data affected classification in MLP. CFS-MLP was able to classify the data using 40 features with 96.90% accuracy.

In the case of prostate data classification, the best classification accuracy reached by CFS feature selection algorithm with linear SVM which is 95.18%, and 64 features required to perform this classification. Normalization and scaling of the prostate data did not have any effect on the required feature number and increased classification accuracies for most of the feature selection algorithms. Entropy as feature selection algorithm using Linear SVM as classifier performed well with 96.18% classification accuracy, but required 100 features to classify for normalized and scaled prostate data. On the other hand, CFS used only 19 features to classify with 95.18% accuracy which is in the acceptable accuracy error range when the trade-off between feature numbers and accuracy is considered. Overall best result for this dataset is reached by MLP using DISR as feature selection algorithm with raw data. DISR-MLP was able to classify normal and cancerous tissues for prostate cancer with 100% classification accuracy and using 80 features/genes. Normalization and scaling affected the classification and required feature number for MLP. DISR-

MLP classified the data with 97% accuracy. Even though the classification accuracy is decreased; the performance is still better and used only 30 features/genes.

For the classification task of DLBCL data, most of the feature selection algorithms performed well when considering the classification accuracies but all of them required more than 50 features to classify. The best classification for raw data performed using CFS as feature selection algorithm with polynomial SVM, 55 features used and classified with 100% accuracy. Normalization and scaling decreased the classification accuracy for all of the feature selection algorithms. When using normalized and scaled data, the best classification rate is reached using Wilcoxon signed-rank test with linear SVM as classifier, 37 features are needed to classify with 94.64% accuracy. On the other hand, CFS-MLP was able to classify with 100% accuracy using 90 features. Bhattacharyya distance as feature selection criterion performed relatively well for raw DLBCL data. MLP was able to classify the data with 95% accuracy. Despite the classification accuracy is decreased, only 50 features used for this performance. Normalization and scaling of the data affected both the classification accuracy and required feature number. MLP was able to classify only using 60 features with 97.32% accuracy.

Microarray datasets studied in this thesis provided corresponding gene descriptions as well. Therefore, the analysis of functions of the selected genes can be analyzed. Feature selection is repeated for raw and normalized data to see the effect of normalization on the process. Feature selection algorithms were stable, and mostly selected the same genes with raw and normalized data.

In [40] Golub et al. selected 50 genes to perform classification, and more than three of the feature selection algorithms that are used in this thesis were able to select 24 genes identical to the original study. Even though no geneticist, biologist or oncologist contributed to the thesis, biological background for commonly selected genes is analyzed. Most of the selected genes play significant roles in cell cycle. However; some of them are not directly related with leukemia. For instance, CD33 molecule is related to the natural kill receptor signaling pathways. Its effect on the apoptosis of AML cells are examined by many studies [78]. Any alteration in this gene may affect the cell cycle and cells may proliferate if they can't perform apoptosis. Even though there are several studies about the relation of the selected

genes, an interdisciplinary and detailed research should be performed for the biological relevancy part of the thesis [79]-[81].

Table 4.74: Most Commonly selected Genes for Leukemia dataset³

Gene Names	
CD33 molecule	Transcription factor 3
Zyxin	Topoisomerase (DNA) II beta
Cystatin C	Granulin
Glutathione S-Transferase, Microsomal	'PRG1 Proteoglycan 1, secretory granule
Amyloid beta precursor-like protein 2	Fumarylacetoacetate hydrolase
Cathepsin D	LYN proto-oncogene
Cyclin D3	CD79a molecule
Complement Factor D (Adipsin)	Myosin light chain 1
Terminal transferase, GSTP1	ATPase H ⁺ transporting V0 subunit c
Myeloperoxidase	Inducible protein
Proteasome Iota Chain	Complement Factor Properdin
Azurocidin	SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily A, Member 4

Additionally, the plot of expression values in Figure 4.61 shows the distinct gene expressions for different types of leukemia and their ability to classify ALL and AML.

³ National Center for Biotechnology Information, 2016
<https://www.ncbi.nlm.nih.gov/>

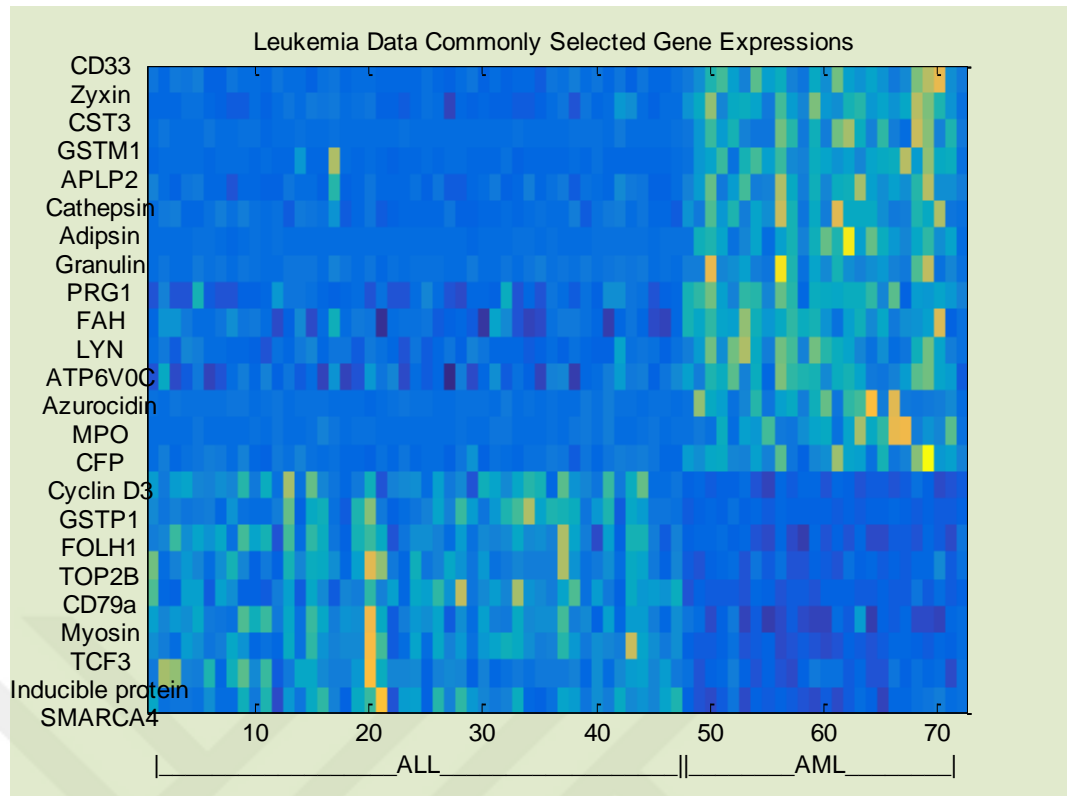


Figure 4.61: Gene Expression Values for 24 commonly selected genes

In [45] Singh et al. could not find any significant biological relation between gene expression values and tumor development. In the same way as Singh's study, most of the selected features/genes are not correlated with the tumor development for prostate cancer. Different feature selection algorithms selected the same 25 genes. Even though there are some reference studies showing relation of these genes to the prostate cancer development, an expanded research should be performed to understand the biological relations of these genes [82]-[89].

Table 4.75: Most Commonly Selected Genes for Prostate Dataset⁴

Gene Names	
Transforming growth factor beta 3	Hepsin
Annexin A2 pseudogene 3	Adipsin
MAF bZIP transcription factor	Prostaglandin D2 synthase
thymosin beta 15a	Neural EGFL like 2
TCR gamma alternate reading frame protein	Crystallin alpha B
PDZ and LIM domain 5	Calmodulin 1
Angiopoietin 1	X-box binding protein 1
ADP ribosylation factor like GTPase 2 binding protein	Serpin family F member 1
Solute carrier family 25 member 6	Ribosomal protein lateral stalk subunit P0
Regulator of G-protein signaling 10	LIM domain only 3
Prolyl 4-hydroxylase subunit beta	Annexin A2
Collagen type IV alpha 6 chain	Family with sequence similarity 107 member A
Latent transforming growth factor beta binding protein 4	

⁴ National Center for Biotechnology Information, 2016
<https://www.ncbi.nlm.nih.gov/>

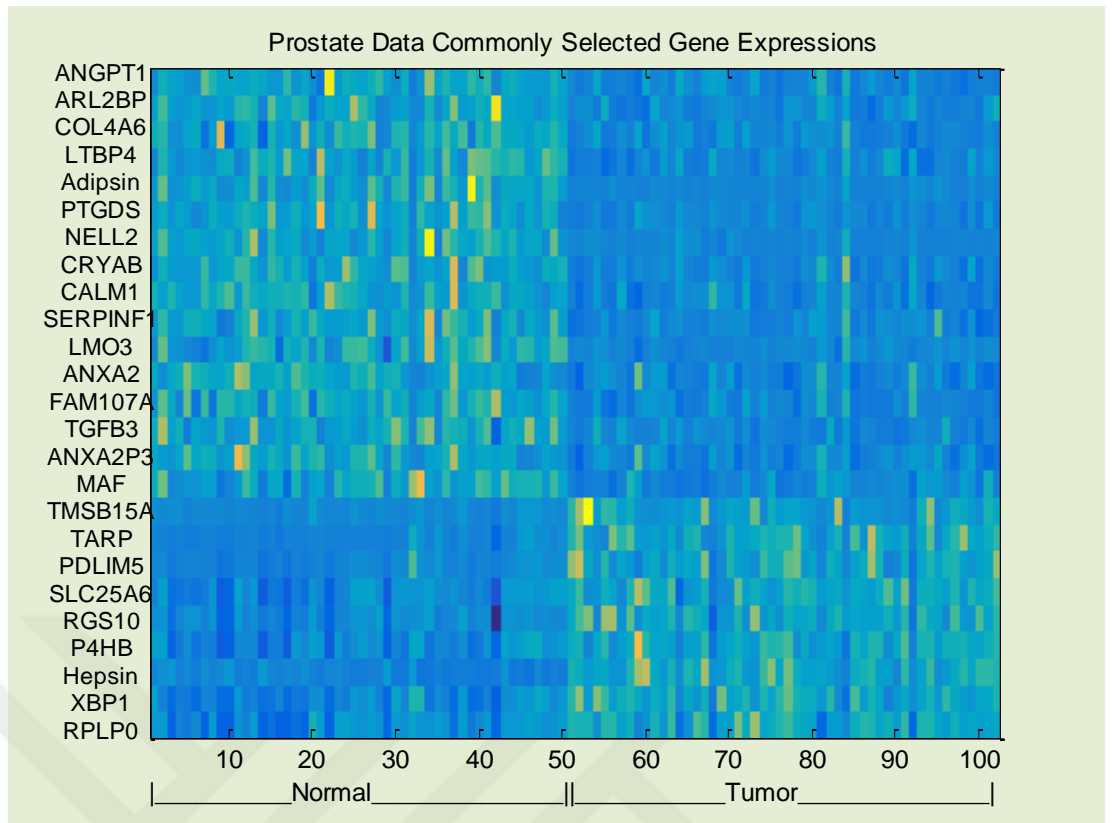


Figure 4.62: Gene Expression Values for 25 commonly selected genes

In the case of DLBCL data, feature selection using Bhattacharyya distance and t test resulted in almost identical gene selection. However; most of the feature selection algorithms resulted in diverse set of gene subsets [47].

Chapter 5

Conclusion

Gene expression data analysis is important for many research areas. It can help the understanding of cancer dynamics and provide great information for biomarker gene selection or targeted drug therapy studies. Even though the biological relevancy is argumentative, it is still preferable comparing to other data types. Yet, gene expression analysis is challenging because of several disadvantages of measured data. The biggest issue when it comes to gene expression data analysis is dimensionality. Generally, gene expression datasets contain values for thousands of genes while remaining small sample size. Therefore, feature selection is crucial for the analysis of gene expression data.

In this thesis, two classification methods are designed by using nine different feature selection algorithms for gene selection and two different classifiers for the classification of cancer types. Discrete data obtained by using microarray data is employed in every method. The basic difference between the proposed methods is caused by the difference in feature selection algorithms. Seven of the feature selection algorithms employed in this thesis rank the features according to their relation to the class and fast in terms of computation time. The other two feature selection algorithms provide a gene subset and computation time is much slower than the others. Therefore, in the first method, data is ranked according to the criterion of feature selection algorithm and top ranked 100 features are selected to perform classification and this process is repeated in each cross-validation. When the feature selection algorithms which provide a gene subset are employed for the gene selection part, a feature subset with 100 features is chosen and the following experiment

procedures are completed over this subset. All experiments are performed for both raw and pre-processed data with or without feature selection. Normalization and scaling is performed in the pre-processing part of the data preparation. Feature selection algorithms were stable enough to choose same features for raw and pre-processed data and pre-processing had no effect on the feature selection. Since SVM is prone to big data, classification without any gene selection is performed only with SVM. Multi-layer perceptrons were not able to classify this type of big data. Feature selection solved the dimensionality problem of the microarray data and made the data available for MLP. Furthermore, different gene subsets for each of the studied cancer types are provided by several feature selection algorithms. Optimality of these gene subsets and the performance of feature selection algorithms are compared in Chapter 5 in detail. SVM with different kernels and MLP with back-propagation algorithm have been employed to classify the cancer types. Performances of the classifiers are examined in Chapter 5 separately for each dataset used in this thesis.

The experiment results in Chapter 5 indicate that both classifiers can be employed for the classification task of microarray cancer datasets. Most of the classifiers perform above 80% classification accuracy. In terms of classification accuracy, every feature selection and classifier combination is applicable to each cancer classification problem. In the case of leukemia type classification, both t test-linear SVM and CFS-MLP combinations resulted in 100% accuracy. The best classification performance is reached by DISR-MLP combination for the classification task of normal and cancerous tissue of prostate cancer. CFS-polynomial SVM and CFS-MLP combinations performed well for DLBCL data set classification problem with 100% percent accuracy. When the biological relevancy of the provided gene subsets are considered, feature selection algorithms fail to select relevant genes for prostate and DLBCL datasets. In the case of leukemia type classification, most of the selected genes by more than three feature selection algorithms were identical to the original study and the classification ability of these genes were distinguishable from Figure 4.61.

In conclusion, proposed feature selection and classification methods performance's resulted better than previous studies in terms of classification ability. It is shown that these methods can be employed for further analysis of gene expression data and biomarker gene selection for specific cancer types. Selected gene

subsets by each of the feature selection method or commonly selected genes by most of the feature selection algorithms were not clinically evaluated or analyzed by a geneticist. Yet, they are compared with the previous study results and some reference studies are found for most commonly selected genes of leukemia and prostate cancer. The classification problems studied in this thesis were binary classification problems. Proposed methods can be enhanced for the multi-class classification problems. Although the results of the experiments in terms of classification accuracy are adequate for most of the feature selection and classifier combinations, optimization of feature selection and classification algorithms must be performed. Optimization of significant parameters may improve the classification accuracy and relativity of the selected genes. Moreover, an interdisciplinary approach must be considered for a better analysis of biological relevancy of the selected gene subsets. This thesis may contribute to these several aspects of gene expression data analysis.

BIBLIOGRAPHY

- [1] Stewart, B., & Wild, C. P. (2016). World cancer report 2014. *World*.
- [2] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., ... & Powell, J. I. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, *403*(6769), 503-511.
- [3] Begum, S., Chakraborty, D., & Sarkar, R. (2016, January). Identifying cancer biomarkers from leukemia data using feature selection and supervised learning. In *2016 IEEE First International Conference on Control, Measurement and Instrumentation (CMI)* (pp. 249-253). IEEE.
- [4] Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., ... & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, *7*(6), 673-679.
- [5] Hu, H., Li, J., Plank, A., Wang, H., & Daggard, G. (2006, November). A comparative study of classification methods for microarray data analysis. In *Proceedings of the fifth Australasian conference on Data mining and analytics-Volume 61* (pp. 33-37). Australian Computer Society, Inc..
- [6] Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, *97*(457), 77-87.
- [7] Lee, J. W., Lee, J. B., Park, M., & Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, *48*(4), 869-885.
- [8] Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., ... & Haussler, D. (2000). Knowledge-based analysis of microarray

- gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1), 262-267.
- [9] Asyali, M. H., Colak, D., Demirkaya, O., & Inan, M. S. (2006). Gene expression profile classification: a review. *Current Bioinformatics*, 1(1), 55-73.
- [10] Reimers, M. (2005). Statistical analysis of microarray data. *Addiction biology*, 10(1), 23-35.
- [11] Chen, Y., & Zhao, Y. (2008). A novel ensemble of classifiers for microarray data classification. *Applied soft computing*, 8(4), 1664-1669.
- [12] Trichopoulos, D., Li, F. P., & Hunter, D. J. (1996). What causes cancer?. *Scientific American*, 275(3), 80-84.
- [13] Camus, M., Siemiatycki, J., & Meek, B. (1998). Nonoccupational exposure to chrysotile asbestos and the risk of lung cancer. *New England Journal of Medicine*, 338(22), 1565-1571.
- [14] Vainio, H., & Boffetta, P. (1994). Mechanisms of the combined effect of asbestos and smoking in the etiology of lung cancer. *Scandinavian journal of work, environment & health*, 235-242.
- [15] Hussein, M. R. (2005). Ultraviolet radiation and skin cancer: molecular mechanisms. *Journal of cutaneous pathology*, 32(3), 191-205.
- [16] Matsumura, Y., & Ananthaswamy, H. N. (2004). Toxic effects of ultraviolet radiation on the skin. *Toxicology and applied pharmacology*, 195(3), 298-308.
- [17] Paruthiyil, S., Parmar, H., Kerekatte, V., Cunha, G. R., Firestone, G. L., & Leitman, D. C. (2004). Estrogen receptor β inhibits human breast cancer cell proliferation and tumor formation by causing a G2 cell cycle arrest. *Cancer research*, 64(1), 423-428.
- [18] STUART, A. (1991). Growth factors and cancer.
- [19] Pike, M. C., Spicer, D. V., Dahmouch, L., & Press, M. F. (1993). Estrogens progestogens normal breast cell proliferation and breast cancer risk. *Epidemiologic reviews*, 15(1), 17-35.
- [20] De Martel, C., Ferlay, J., Franceschi, S., Vignat, J., Bray, F., Forman, D., & Plummer, M. (2012). Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *The lancet oncology*, 13(6), 607-615.

- [21] zur Hausen, H. (2009). The search for infectious causes of human cancers: where and why. *Virology*, 392(1), 1-10.
- [22] Giovannucci, E. (2001). Insulin, insulin-like growth factors and colon cancer: a review of the evidence. *The Journal of nutrition*, 131(11), 3109S-3120S.
- [23] Jee, S. H., Kim, H. J., & Lee, J. (2005). Obesity, insulin resistance and cancer risk. *Yonsei medical journal*, 46(4), 449-455.
- [24] Seely, S., & Horrobin, D. F. (1983). Diet and breast cancer: the possible connection with sugar consumption. *Medical hypotheses*, 11(3), 319-327.
- [25] Rountree, M. R., Bachman, K. E., Herman, J. G., & Baylin, S. B. (2001). DNA methylation, chromatin inheritance, and cancer. *Oncogene*, 20(24), 3156-3165.
- [26] Jasperson, K. W., Tuohy, T. M., Neklason, D. W., & Burt, R. W. (2010). Hereditary and familial colon cancer. *Gastroenterology*, 138(6), 2044-2058.
- [27] Hemminki, K., & Li, X. (2004). Familial risks of cancer as a guide to gene identification and mode of inheritance. *International journal of cancer*, 110(2), 291-294.
- [28] Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., ... & Kaasa, S. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *Journal of the national cancer institute*, 85(5), 365-376.
- [29] Baskar, R., Lee, K. A., Yeo, R., & Yeoh, K. W. (2012). Cancer and radiation therapy: current advances and future directions. *Int J Med Sci*, 9(3), 193-199.
- [30] Pucci, B., Kasten, M., & Giordano, A. (2000). Cell cycle and apoptosis. *Neoplasia*, 2(4), 291-299.
- [31] Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *cell*, 100(1), 57-70.
- [32] Roth, C. M. (2002). Quantifying gene expression. *Current issues in molecular biology*, 4, 93-100.
- [33] Mikkilineni, V., Mitra, R. D., Merritt, J., DiTonno, J. R., Church, G. M., Ogunnaike, B., & Edwards, J. S. (2004). Digital quantitative measurements of gene expression. *Biotechnology and bioengineering*, 86(2), 117-124.

- [34] Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270(5235), 484.
- [35] Babu, M. M. (2004). Introduction to microarray data analysis. *Computational genomics: Theory and application*, 225-249.
- [36] Piatetsky-Shapiro, G., & Tamayo, P. (2003). Microarray data mining: facing the challenges. *ACM SIGKDD Explorations Newsletter*, 5(2), 1-5.
- [37] Schulze, A., & Downward, J. (2001). Navigating gene expression using microarrays—a technology review. *Nature cell biology*, 3(8), E190-E195.
- [38] Butte, A. (2002). The use and analysis of microarray data. *Nature reviews drug discovery*, 1(12), 951-960.
- [39] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., ... & Gaasterland, T. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature genetics*, 29(4), 365-371.
- [40] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439), 531-537.
- [41] Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive evaluation of multiclassification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5), 631-643.
- [42] Lee, J. W., Lee, J. B., Park, M., & Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4), 869-885.
- [43] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 906-914.
- [44] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3), 389-422.

- [45] Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., ... & Lander, E. S. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2), 203-209.
- [46] Peng, Y. (2006). A novel ensemble machine learning for robust microarray data classification. *Computers in Biology and Medicine*, 36(6), 553-573.
- [47] Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., ... & Ray, T. S. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8(1), 68-74.
- [48] Chopra, P., Lee, J., Kang, J., & Lee, S. (2010). Improving cancer classification accuracy using gene pairs. *PLoS One*, 5(12), e14305.
- [49] Yang, K., Cai, Z., Li, J., & Lin, G. (2006). A stable gene selection in microarray data analysis. *BMC bioinformatics*, 7(1), 1.
- [50] Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1), 4-37.
- [51] Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., & Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11), 1454-1461.
- [52] Breitling, R., & Herzyk, P. (2005). Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *Journal of bioinformatics and computational biology*, 3(05), 1171-1189.
- [53] Li, J., & Fine, J. P. (2008). ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics*, 9(3), 566-576.
- [54] Mamitsuka, H. (2006). Selecting features in microarray classification using ROC curves. *Pattern Recognition*, 39(12), 2393-2404.
- [55] Ma, S., & Huang, J. (2005). Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics*, 21(24), 4356-4362.
- [56] Zhang, J. G., & Deng, H. W. (2007). Gene selection for classification of microarray data based on the Bayes error. *BMC bioinformatics*, 8(1), 370.

- [57] Mei, R., Di, X., Ryder, T. B., Hubbell, E., Dee, S., Webster, T. A., ... & Smeekens, S. P. (2002). Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, 18(12), 1593-1599.
- [58] Liu, X., Krishnan, A., & Mondry, A. (2005). An entropy-based gene selection method for cancer classification using microarray data. *BMC bioinformatics*, 6(1), 1.
- [59] Yan, X., Deng, M., Fung, W. K., & Qian, M. (2005). Detecting differentially expressed genes by relative entropy. *Journal of theoretical biology*, 234(3), 395-402.
- [60] Kira, K., & Rendell, L. A. (1992, July). The feature selection problem: Traditional methods and a new algorithm. In *AAAI* (Vol. 2, pp. 129-134).
- [61] Kira, K., & Rendell, L. A. (1992, July). A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning* (pp. 249-256).
- [62] Kononenko, I., Šimec, E., & Robnik-Šikonja, M. (1997). Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, 7(1), 39-55.
- [63] Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F., & Mewes, H. W. (2005). Gene selection from microarray data for cancer classification—a machine learning approach. *Computational biology and chemistry*, 29(1), 37-46.
- [64] Wang, Y., & Makedon, F. (2004, August). Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data. In *Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE* (pp. 497-498). IEEE.
- [65] Hall, M. A. (1999). *Correlation-based feature selection for machine learning* (Doctoral dissertation, The University of Waikato).
- [66] Meyer, P. E., Schretter, C., & Bontempi, G. (2008). Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, 2(3), 261-274.
- [67] Bontempi, G., & Meyer, P. E. (2010). Causal filter selection in microarray data. In *Proceedings of the 27th international conference on machine learning (icml-10)* (pp. 95-102).

- [68] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8), 1226-1238.
- [69] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
- [70] Chen, P. H., Lin, C. J., & Schölkopf, B. (2005). A tutorial on v-support vector machines. *Applied Stochastic Models in Business and Industry*, 21(2), 111-136.
- [71] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.
- [72] Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1), 1.
- [73] Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Ares, M., & Haussler, D. (1999). Support vector machine classification of microarray gene expression data. *University of California, Santa Cruz, Technical Report UCSC-CRL-99-09*.
- [74] Wang, S. C. (2003). Artificial neural network. In *Interdisciplinary Computing in Java Programming* (pp. 81-100). Springer US.
- [75] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14), 2627-2636.
- [76] Atiya, A. (1991). *Learning algorithms for neural networks* (Doctoral dissertation, California Institute of Technology).
- [77] Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.
- [78] Malik, M., Chiles, J., Xi, H. S., Medway, C., Simpson, J., Potluri, S., ... & Crane, P. (2015). Genetics of CD33 in Alzheimer's disease and acute myeloid leukemia. *Human molecular genetics*, ddv092.
- [79] Weng, H. Y., Huang, H. L., Zhao, P. P., Zhou, H., & Qu, L. H. (2012). Translational repression of cyclin D3 by a stable G-quadruplex in its 5' UTR: implications for cell cycle regulation. *RNA biology*, 9(8), 1099-1109.

- [80] Kozlov, I., Beason, K., Yu, C., & Hughson, M. (2005). CD79a expression in acute myeloid leukemia t (8; 21) and the importance of cytogenetics in the diagnosis of leukemias with immunophenotypic ambiguity. *Cancer genetics and cytogenetics*, 163(1), 62-67.
- [81] Yuan, N., Song, L., Lin, W., Cao, Y., Xu, F., Liu, S., ... & Zhang, H. (2015). Autophagy collaborates with ubiquitination to downregulate oncoprotein E2A/Pbx1 in B-cell acute lymphoblastic leukemia. *Blood cancer journal*, 5(1), e274.
- [82] Kim, H. J., Han, J. H., Chang, I. H., Kim, W., & Myung, S. C. (2012). Variants in the HEPsin gene are associated with susceptibility to prostate cancer. *Prostate cancer and prostatic diseases*, 15(4), 353-358.
- [83] Kim, D. H., Roh, Y. G., Lee, H. H., Lee, S. Y., Kim, S. I., Lee, B. J., & Leem, S. H. (2013). The E2F1 oncogene transcriptionally regulates NELL2 in cancer cells. *DNA and cell biology*, 32(9), 517-523.
- [84] Bao, L., Loda, M., Janmey, P. A., Stewart, R., Anand-Apte, B., & Zetter, B. R. (1996). Thymosin beta 15: a novel regulator of tumor cell motility upregulated in metastatic prostate cancer. *Nature medicine*, 2(12), 1322-1328.
- [85] Fritzsche, F. R., Stephan, C., Gerhardt, J., Lein, M., Hofmann, I., Jung, K., ... & Kristiansen, G. (2010). Diagnostic and prognostic value of T-cell receptor gamma alternative reading frame protein (TARP) expression in prostate cancer. *Histology and histopathology*, 25(4), 733.
- [86] Zhao, J., Chen, L., Shu, B., Tang, J., Zhang, L., Xie, J., ... & Qi, S. (2015). Angiopoietin-1 Protects the Endothelial Cells Against Advanced Glycation End Product Injury by Strengthening Cell Junctions and Inhibiting Cell Apoptosis. *Journal of cellular physiology*, 230(8), 1895-1905.
- [87] Zamora, M., Granell, M., Mampel, T., & Viñas, O. (2004). Adenine nucleotide translocase 3 (ANT3) overexpression induces apoptosis in cultured cells. *FEBS letters*, 563(1-3), 155-160.
- [88] Mundel, T. M., Yliniemi, A. M., Maeshima, Y., Sugimoto, H., Kieran, M., & Kalluri, R. (2008). Type IV collagen $\alpha 6$ chain-derived noncollagenous domain 1 ($\alpha 6$ (IV) NC1) inhibits angiogenesis and tumor growth. *International Journal of Cancer*, 122(8), 1738-1744.

[89] Varisli, L. (2013). Identification of new genes downregulated in prostate cancer and investigation of their effects on prognosis. *Genetic testing and molecular biomarkers*, 17(7), 562-566.

